The Changing Patterns of Payments in the United States

Based on a speech delivered by President Santomero at the 25th SUERF Colloquium: "Competition and Profitability in European Financial Services: Strategic, Systemic, and Policy Issues," Marjolin Lecture, Madrid, Spain, October 16, 2004

BY ANTHONY M. SANTOMERO

Although the origins and evolution of payment structures in the United States and Europe are quite different, both systems are moving toward more electronic forms of payment. In "The Changing Patterns of Payments in the United States," President Santomero highlights the differences between U.S. and European payments infrastructure; discusses how the roots and evolution of the U.S. payments system differs from Europe's; and outlines the likely path of the U.S. payment system and the Fed's role in it.

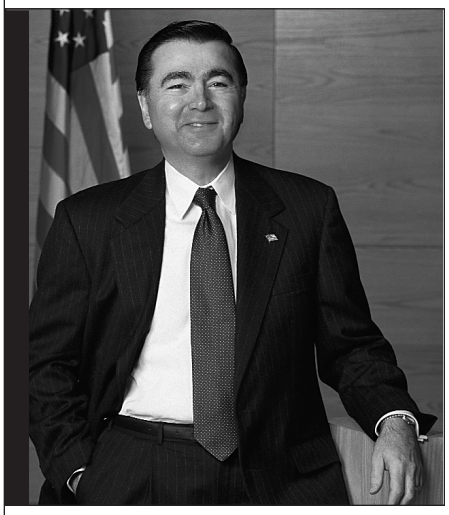
As a career academic and current U.S. central banker, I would like to offer commentary on some changes taking place in the financial services industry in the United States. Specifically, I would like to discuss what is happening in the U.S. payments system. The changes occurring in the U.S. are interesting in their own right and as a point of comparison and contrast with what is happening in the European payments arena.

As anyone who knows the sector would readily admit, the origins and evolution of payment structures in the United States and Europe could not be more different. Now, however, we are beginning to see signs that the two systems are starting to converge. Both are moving toward more electronic payment services through a number of vehicles. In other words, two systems that started out quite differently are converging toward similar systems. On the U.S. side, the pattern of payments is indeed evolving — some might say it is experiencing a radical change. America's paper-based payments system is giving way to a new realm of electronic payments vehicles — a transition that has already occurred in Europe. Indeed, there has been quite a bit of diversity in the forms of payments used in the U.S. However, as is typical in this area, change has been, and will be, greatly affected by our financial history and its legacy systems.

This presents the Federal Reserve System with many challenges because, unlike most central banks in Europe, the Federal Reserve is not only a regulator but also a service provider. It has been a vital part of the retail payments system since its founding more than 90 years ago. From its inception, the Federal Reserve has had a dual role as the central bank charged with ensuring the integrity of the payments system and as a participant in its evolution.

Over time, the Fed's role in payments and that of European central banks are likely to converge as well. The Fed's role in paper processing will likely diminish over time as checks recede in both absolute volume and relative importance in our retail payments system. As this occurs, it will further our resemblance to the central banks of Europe. Over time, both the Fed and European central banks will concentrate more of their efforts on their services on large-dollar gross settlement, with TARGET2 likely following the evolution of Fedwire.[1]

[1] TARGET is the system used in the countries of the European Union for the settlement of central bank operations, large-value euro interbank transfers, and other euro payments. TARGET2, the next generation of the system, is currently under development.



Anthony M. Santomero, President, Federal Reserve Bank of Philadelphia

With that prologue, I would like to share my thoughts on payments, concentrating on three issues:

- the current status of the U.S. payments infrastructure vis-à-vis Europe's;
- how the roots and evolution of the U.S. payment system differ from those of Europe; and
- the likely future path of the U.S. payment system and the Fed's role in it, with an emphasis on how we are likely to resemble Europe and how we will be different.

THE CURRENT STATE OF PAYMENTS TECHNOLOGY IN THE U.S.

Historically, Americans and Europeans have long relied on an entirely different mix of payments vehicles. For example, Europeans use cash roughly twice as much as Americans. However, looking at noncash transactions gives evidence of where the differences truly lie. In Europe, half of all noncash retail payments are made through a Giro system and only about 15 percent are made by check. In the United States, it is almost exactly the reverse. Half of all noncash retail payments are made by paper check and less than 10 percent are made through ACH, which is the American version of a Giro system.[2]

The dominance of the Giro in Europe and of the check in the United States is a long-standing feature of our respective payment systems. The history of how this dominance evolved is interesting and instructive, as I will elaborate later.

Payment cards account for the remainder of retail payments, and there are similarities and differences between Europe and the United States. The similarities lie in the use of debit cards. Debit cards, a relatively recent innovation, have caught on quickly both in Europe and in the U.S., and they now account for about a quarter of noncash retail payments in both places. The differences are in our use of credit cards. Credit cards have long been an important payment vehicle in the U.S. and, at present, account for about a quarter of our noncash retail payments. In Europe, credit cards are used less frequently — in less than 10 percent of transactions, though I would note that Europeans' use of credit cards has picked up in recent years.[3]

Credit cards have long been an important payment vehicle in the U.S. and, at present, account for about a quarter of our noncash retail payments.

The long-standing success of the credit card in the U.S. and the rapid rise of the debit card in both Europe and the U.S. are also interesting and instructive stories, which I will touch on as well. First, let's begin with the story of the Giro and the check.

The European Structure. To understand the dominance of the Giro in Europe and the check in the U.S. we have to go back about 100 years to the late 19th and early 20th centuries. At that time, European banks did not provide routine payment services. They served primarily as merchant banks and as private banks for wealthy individuals.

In the late 1800s, local post offices began establishing postal Giro systems as a convenient way for common people to deposit savings, and these systems later evolved to allow people to remit and receive payments. The system was successful in that it allowed every post office savings account holder to make and receive payments both locally and nationally. This revolutionary achievement rendered noncash payment transactions accessible to large sectors of the population.

Later, in the 1950s and 1960s, European banks sought to broaden their business lines to encompass the mass market as a way to expand their deposit base to fund loans. This meant providing routine payment services to customers; so bank Giro systems were created to handle the volume.

This evolution occurred relatively smoothly and rapidly as a result of Europe's concentrated banking industry — a few banks operating nationwide, cooperating closely with each other.

At the same time, European governments wanted to establish payment systems that minimized costs and maximized access. The advent of technological advances created such opportunities through electronification. When technology made it economical to replace paper Giros with electronic Giros, European governments pushed

[2] Data from Bank for International Settlements, cited in *Statistics on Payment and Settlement Systems in Selected Countries*, March 2004 (figures for 2002), prepared by the Committee on Payment and Settlement Systems of the Group of 10 Countries.

[3] Data from Bank for International Settlements, cited in *Statistics on Payment and Settlement Systems in Selected Countries*, March 2004 (figures for 2002), prepared by the Committee on Payment and Settlement Systems of the Group of 10 Countries.

for the transition, and the concentration of the payments system in the hands of the postal service and a few national banks made it relatively easy to accomplish. Because of its Giro system, Europe had, or could easily set up, centralized accounts for credit transfers. In short, European central banks encouraged — and in some cases mandated — the use of electronic Giro systems.

The U.S. Structure. In contrast, the U.S. payments system evolved quite differently from Europe's. Historically, U.S. banks tended to provide services, including payment services, to the broad spectrum of people and businesses. On the loan side, commercial banks focused on commercial and industrial lending, but they took deposit balances from all economic strata.

In early America, the geographical expanse of the country encouraged a fragmented system wherein state banks issued their own notes. Entry into the banking business was relatively easy, but bank branching was very restricted. Banks were prohibited from branching outside their home state, and in many states, branching was restricted still further. As a consequence, a region would be served by a relatively large number of banks, but there were no banks operating nationwide.

To effect transactions, people paid one another with paper checks drawn on their bank or paper currency notes issued by their bank. The banks would then clear these checks and notes among themselves.

With so many individual banks spread out across such a big country, and banks clearing paper instruments among themselves, effecting transactions outside the local area was cumbersome. When someone received a bank check or a bank note as payment and deposited it at his bank, the bank would discount the instrument's value based on the cost of presenting it to the "drawn on" bank for payment and some assessment of the creditworthiness of the "drawn on" bank. The farther away the bank, the less familiar its financial condition and the greater the transportation cost associated with clearing the instrument, and so the greater the discount tended to be. So a merchant in Kansas City, Missouri, accepting as payment a check drawn on a bank in Allentown, Pennsylvania, knew he would be credited with less than the face or par value of the check and would have to consult with his bank to find out how much less. Obviously, this was a payment system inimical to the growth of national commerce.

By the turn of the 20th century, it was clear that the U.S. needed a more well-integrated national payment system. Indeed, one of the main reasons Congress established the Federal Reserve System in 1913 was to create a national clearing system in which checks could exchange at par value. To achieve this, the Federal Reserve offered check-clearing services free of charge to banks that joined the Fed System.

However, the Fed did not become the sole provider of check-clearing services, despite offering its services for free. First, not all banks chose to join the Fed System, primarily because of some of the regulatory implications. In addition, large correspondent banks offered smaller respondent banks an array of "bankers' bank services," including check clearing, and banks could take advantage of local and national clearinghouse arrangements.

Nonetheless, the Fed established a large market presence, providing a baseline level of national check-clearing services accessible to all banks, large and small, anywhere in the country. Thus, the Fed contributed to the viability of both the paper check and the small community bank.

In the 1960s and 1970s, U.S. banks and the Fed applied advances in computing technology to check processing, increasing the efficiency of their operations. Banks found the paper check payments business to be profitable, and consumers were quite comfortable and confident in the use of checks.

In short, checks were the dominant form of noncash payment, and there was little momentum for change in the U.S. payments system. One might argue that bank Giro systems, which were arising in Europe at the time, would have increased the efficiency of the payments system even more. Yet with so many banks in the U.S. — all serving local markets

— developing the legal framework, industry standards, and institutional arrangements necessary to establish such a payments network nationally would have been a daunting task. In any case, American banks are forbidden under antitrust law to work together.

The Fed itself introduced its version of an electronic Giro system in the early 1970s. We call it the automated clearinghouse, or Fed ACH. Fed ACH has met with some success.

However, unlike the European Giro, ACH has not developed into the dominant form of electronic payment, in part, because, traditionally, only banks — not individuals — could initiate ACH payments. This made ACH practical only for companies engaged in batch-processing a large number of payments, such as payroll disbursement.

In a typical transaction, a firm would forward to its bank an electronic file containing payments to be made from the firm's account. The bank would then initiate the ACH transactions by sending the file to the Fed, which would transfer funds from the bank's account to the accounts of the various payees' banks, and then notify them of the account holders to be credited.

I will add that a relatively recent variant allows large organizations to collect regular payments using the ACH. A typical transaction of this nature would involve individual customers' authorizing their bank to make ACH payments directly to a firm — perhaps their utility company or mortgage company — on a recurring basis.

CARDS DRIVE CHANGES IN U.S. PAYMENTS

While Fed ACH saw some success as a means to effect electronic payments, it was the credit card that proved most instrumental in moving U.S. payments from paper to electronics. The credit card was the first electronic payments instrument to emerge in the U.S. Credit cards were introduced in the 1950s, and their use grew rapidly over the next three decades.

Credit Cards. Not coincidentally, the U.S. credit card infrastructure looks a lot like the European banking system. There are relatively few major card associations; they operate nationwide; and they are not subject to the anti-trust laws that prohibited collaboration among U.S. banks. In fact, the credit card associations benefited from some early antitrust rulings against banks.

While Fed ACH saw some success as a means to effect electronic payments, it was the credit card that proved most instrumental in moving U.S. payments from paper to electronics.

In the 1990s, when the tech boom made information processing and telecommunications more powerful and less expensive, the credit card associations were well positioned to take full advantage of these developments. Low-cost telecom has made real-time, point-of-service verification of cardholders and their credit status widespread, speeding transactions and curtailing fraud. Of significance for the future, this technology has made the credit card a viable means of payment for e-commerce.

Debit Cards. After the credit card, the debit card is the second most popular electronic instrument for making retail payments in the U.S. today. The debit card arrived on the scene relatively recently — during the 1980s — in both the United States and Europe. But since its arrival, growth in usage has been dramatic.

In Europe, the debit card emerged as an evolution of banks' automated teller machine (ATM) systems. Instead of using their card to withdraw cash from an ATM to pay merchants, bank customers simply present their card to the merchants, and their bank account is debited directly.

This same progression occurred in the U.S. But in the U.S., the credit card networks responded with debit card products of their own. Visa and MasterCard already had an infrastructure for processing credit card transactions at the point of sale. They leveraged this infrastructure to establish offline debit card networks. Indeed, in the U.S., these so-called "signature" debit cards are proving at least as popular as ATM, or "PIN-based," debit cards.[4]

Signature debit cards now account for about two-thirds of the total of debit transactions. So it could be said that they are even more popular than their PIN counterparts. However, PIN-based debits are growing a bit faster than signature.[5]

In any case, debit cards, in general, seem to be leading the migration away from cash and checks and toward electronic payments in the U.S. This trend is substantiated by the Survey of Consumer Finances, sponsored by the Federal Reserve Board of Governors

[4] See the conference summary "Prepaid Cards: How Do They Function? How Are They Regulated?" produced by the Federal Reserve Bank of Philadelphia's Payment Cards Center, June 2004, available at: www.philadelphiafed.org/pcc/conferences/PrepaidCards_062004.pdf.

[5] See the Retail Payments Research Project: *A Snapshot of the U.S. Payments Landscape*, Federal Reserve System, 2002.

and compiled by the Research Department at the Philadelphia Fed.[6]

The survey indicates that less than 18 percent of households used debit cards in 1995. By 2001, nearly half of all households were using them. Not coincidentally, the survey also divulged a substantial reduction in the use of cash over the same period.[7]

The growing popularity of debit cards in the U.S. seems to be part of a broader phenomenon. As I mentioned earlier, debit cards have caught on just as quickly in Europe. In fact, recently, for the first time ever, Visa's global debit sales volume surpassed its credit sales volume.[8]

THE FUTURE OF THE U.S. RETAIL PAYMENTS SYSTEM

By now, I hope I have given you some perspective on the current state of U.S. retail payments and the evolutionary process that brought us there.

Looking ahead, retail payments in the U.S. will continue moving away from cash and paper checks and toward electronic instruments, including credit cards, debit cards, ACH, and emerging vehicles such as prepaid cards.

Though roughly half of our noncash payments are still being made by paper check, the tide has turned. In fact, recent research by the Federal Reserve shows check usage peaked in the mid-1990s and has been declining steadily ever since. So paper checks are not only losing market share, they are actually declining in volume and have been for about a decade.[9]

The share of retail transactions handled by cards will continue to grow in the U.S., particularly at the point of sale. Debit cards have made particularly deep inroads in the realm of "micropayments" — purchases under $20. According to a survey by MasterCard International, debit cards now account for about one-third of all micropayments, a 61 percent increase over 2001.[10] Visa claims to have authorized 82 percent more payments at quick-service restaurants between January and July of 2004 than during the same period in 2003.[11] Here we see debit transactions replacing cash, since the survey indicated a substantial drop in cash micropayments.

In the future, organizations other than banks will expand their role in the payments system, especially retailers themselves.

Several fast-food chains are promoting greater use of payment cards at their restaurants. (It undoubtedly has not escaped their attention that customers spend, on average, over 50 percent more when they pay with a card rather than cash.[12]) This movement has tremendous upside potential. Last year, consumers used their cards to spend $6.5 billion at fast-food restaurants, and that was with only 10 percent of such restaurants accepting cards.[13]

In the future, organizations other than banks will expand their role in the payments system, especially retailers themselves. As a result of recent legal action brought by Wal-Mart against U.S. card companies, retailers now appreciate the costs and benefits associated with alternative payment processing arrangements and will weigh in to protect their interests. As you may know, Wal-Mart, the largest retailer in the U.S., along with other merchants, balked at the idea of accepting signature debit cards — and their associated fees — without the right to negotiation. They sued U.S. bank credit card associations, prevailing in a good portion of their efforts. Their settlement eliminated the "honor all cards" rule, effectively allowing merchants to decline signature debit products without jeopardizing their ability to accept credit products or PIN debit cards.

In short, I expect keen competition among card providers and aggressive marketing by both card providers and merchants to increase the speed with which cards replace paper for point-of-sale transactions in the U.S.

How quickly U.S. consumers move from paper to electronics, when it comes to bill paying, is an interesting question. The speed and scope of that transition depend on the evolution of our payments system.

---

[6] See Loretta J. Mester, "Changes in the Use of Electronic Means of Payment: 1995-2001," Federal Reserve Bank of Philadelphia *Business Review*, Third Quarter 2003.

[7] See Mester, Federal Reserve Bank of Philadelphia *Business Review*, Third Quarter 2003.

[8] Press release, "Visa Global Debit Card Sales Volume Surpasses Credit," Visa International, April 20, 2004.

[9] See the Federal Reserve System Retail Payments Study, December 2004.

[10] David Breitkopf, "MasterCard, Pulse Report Wider Use of Debit Cards," *American Banker*, May 17, 2004.

[11] W.A. Lee, "CEO Confident as Visa Posts More Records," *American Banker*, August 5, 2004.

[12] Data from W.A. Lee, "CEO Confident as Visa Posts More Records," *American Banker*, August 5, 2004.

[13] "Cards…at participating restaurants," *Electronic Payments International*, August 19, 2004.

As I mentioned earlier, the ACH system in the U.S. has not been as successful as Europe's Giro systems. But things may be changing. Financial institutions are finding innovative new uses for ACH, spanning a broad range of retail transactions and shifting substantial volumes to this system, primarily at the expense of check volume.

The most important of these innovations is accounts receivable check (ARC) conversion. Large organizations that receive paper checks from customers as remittance for retail payments are now scanning the checks to digitally capture their relevant payment information. The companies can then use this information to create an electronic file, which is then transmitted to an ACH payments provider — usually the Fed — for processing. In some cases, even individual merchants who accept customer checks at the point-of-sale can use the information on the check to generate an electronic file. That file is then sent to the merchant's bank for processing through the ACH.

Conversion to ACH is helping to streamline payments initiated by check, even when the paper check would follow. It is also being used to process one-time payments initiated via the Internet.

As the owner/operator of the Fed ACH system, the Federal Reserve has been working to ensure its ACH system is equipped to accommodate changes in volumes and the nature of payments, even as these applications proliferate. As in check processing, the Fed is not the sole provider of ACH. Though the Federal Reserve network currently originates about two-thirds of all ACH payments volume, we are also seeing growth among private-sector ACH networks. Indeed, as ACH continues to gain acceptance as a payment vehicle, its products and marketing will evolve so as to make it more attractive and accessible to individuals and businesses.

MANAGING THE TRANSITION

So the private sector is shifting retail payments in the U.S. away from paper-based instruments and toward electronic ones. But history tells us that people's payment habits change only gradually. When people are comfortable with, and confident in, a payment structure, they are reluctant to give it up. As a result, the paper check is likely to be with us for some time.

In the meantime, the Fed has been trying to take full advantage of the efficiencies afforded by electronic processing of payments initiated by paper check in the interest of maximizing the efficiency of the payment system. Thus, the Fed is doing what it can to foster check truncation and electronification at as early a stage as possible in the payment process.

The Fed is now well positioned to pursue this objective. Two pieces of legislation have set the stage. One is a law that has been on the books for 25 years: the Monetary Control Act of 1980. The second was passed in 2003 and went into effect in October 2004: the Check Clearing for the 21st Century Act, commonly called Check 21. Let me explain the significance of each.

Recall that when the Fed began its check processing operations, it provided the service at no charge to its member banks. The Monetary Control Act of 1980 changed all that. It required the Fed to offer its payment services to all banks at prices fully reflecting the Fed's costs of production, including imputed profits. This change established a marketplace incentive for the Fed and its private-sector competitors in check processing to maximize the efficiency of their check processing operations.

The Fed has been trying to take full advantage of the efficiencies afforded by electronic processing of payments initiated by paper check in the interest of maximizing the efficiency of the payment system.

The second piece of legislation, Check 21, adds an important new dimension to the competitive drive for greater efficiency in check processing. The essence of the new law is that it makes the facsimile of a check created from an electronic image serve as the legal equivalent of the check itself. In doing so, it eliminates a significant legal barrier to check truncation and electronification of check processing. A collecting bank can soon create an electronic image of a check, transmit the image to the paying bank's location, and then present the paying bank with a paper reproduction or with the electronic image itself. The hope and expectation is that gradually more and more paying banks will prefer the image itself.

Accepting images for both deposit and presentment eliminates back office capture of the check as well as the inconvenience of physical transportation. Indeed, under the new Check 21 legislation, it will become even easier to move toward a more electronic check process because banks will be provided with additional

options for processing image-based payments.

As a provider of financial services, the Fed has been actively engaged in bringing a whole array of image products to market to take advantage of the capability of image clearing. The Fed has established an image archive for electronic items; it has enhanced the ability to produce facsimile checks; and it has extended clearing times to encourage the use of the new image technology that the act allows. In short, the Fed is introducing new services that will enable banks to take full advantage of Check 21.

How fast will the transition occur? Our best guess is that the industry will be slow to embrace the new capabilities that the law permits. We must also consider the possibility that making check processing more efficient will actually extend the life of the waning check. In any case, the Federal Reserve Banks' financial services division is committed to working with the industry to ensure a smooth transition.

THE CHALLENGE TO THE FED

With the evolution of the payments system in the U.S. accelerating, the Federal Reserve must make some major adjustments to its payments services as the changing payments system alters its role. Nonetheless, the Fed is committed to working to improve the reliability and efficiency of the current generation of payments vehicles, even as it works to foster innovation and to support the next generation of payments vehicles. Both commitments are equally important during this period of transition.

With this dual commitment in mind, the Fed continues to fulfill its traditional role as payments processor even while it supports the move to the new electronic clearing environment. Striking the right balance between these two seemingly divergent goals is a challenge. Nonetheless, the Fed has begun implementing a strategy that includes key elements to help it successfully meet both commitments.

The Fed has recently announced a program of "aggressive electronification" of retail payments in the U.S. This push toward electronics will help facilitate Check 21 and quicken the transition to an all-electronic world. The Fed is also investing heavily in technologies that enable electronification. In addition, as check volumes decline, the pressure has been on to find new processing efficiencies. The transition will not be easy, particularly for the Federal Reserve System.

By setting prices that reflect the low cost of electronic check processing relative to paper, the Fed will allow, indeed encourage, the market to drive checks toward electronics.

The Fed currently clears about one-third of all checks written in the U.S. As check volumes have declined, the Fed has had to consolidate its operations, closing down processing sites where appropriate. Nonetheless, it has attempted to maintain reasonable service levels nationally by re-routing checks to nearby sites.

So that you can see the scale of this effort, I will note that two years ago the Fed had 45 check processing sites. By the end of 2006, we expect to be down to 22. This downsizing to match costs and revenues helps the Fed fulfill its traditional role of payments processor while at the same time maintaining efficiency in this new environment.

Such a radical transformation within the Fed's financial services division is made necessary by law. As I mentioned, the Monetary Control Act mandated that the Fed set prices on its services to fully recover its costs. At the same time, the Fed is required to adjust its portfolio of services to correspond to the clearing needs of the industry. As such, the aggregate decline in volume in this volume-based service creates a substantial challenge to the System. Achieving full cost recovery will become more challenging for the Fed as the volume of check usage continues to decline.

Nonetheless, by setting prices that reflect the low cost of electronic check processing relative to paper, the Fed will allow, indeed encourage, the market to drive checks toward electronics. In addition, the Fed will continue to develop its capabilities and expand its electronics capacity to respond to the market's evolution and consumers' needs. The impact of these changes and those that follow will ultimately transform the U.S. payments system and enable a radical restructuring of its service capabilities.

A WORD ABOUT WHOLESALE PAYMENTS

Before closing, let me briefly discuss the Fed's wholesale payments operation. Aside from its role in supporting retail payments, or small-dollar transactions, the Fed has long had a role in facilitating wholesale, or large-dollar, transactions. Fedwire is the Fed's real-time wholesale payments operation used to transfer both funds and securities. Fedwire transactions typically involve large-value, time-critical payments, such as payments for the settlement of interbank purchases and

sales of federal funds, or securities or real estate transactions.

Fedwire first went into operation back in 1918, and its operations have evolved with advances in technology and the integration of financial markets. The Fed has recently centralized Fedwire operations from all 12 Reserve Banks to its New York Bank — with both a hot and a cold backup.

Now, a parallel process seems to be in motion in Europe. The initiative known as TARGET2 will likely consolidate European central banks' wire transfer operations. As in the case of Fedwire, this standardized processing platform will reduce costs through economies of scale and improve flexibility of wholesale payments.

CONCLUSION

My purpose here was to review and explain the state of payments technology in the U.S. vis-à-vis that of Europe. The roots of these two payment systems lie in the different banking structures of the U.S. and Europe and different perceptions of appropriate regulation.

Europe's is a system of a few large banks that can easily be regulated into a centralized world — first with near-universal Giro accounts and soon with an electronic world of more centralized clearing.

In the U.S., markets and consumers led the nation to a multiplicity of banks and a payments system that has been paper intensive. This is changing in the U.S., as cards replace checks, and electronic clearing truncates the maze of paper that fills U.S. post offices. Indeed, it seems the U.S. payments system is moving toward convergence with the European model. Our progress, while promising, occurs largely in fits and starts. The U.S. is a large nation with many providers, much complexity, and a philosophy of market-based solutions.

This has presented challenges for the Federal Reserve as a provider of financial services. It has necessitated restructurings, plant closings, and difficult decisions that most central banks in Europe have been spared. Yet, by law, the Fed is charged with the dual role of a regulator seeking to maintain the stability and efficiency of the payments system and a provider of payment services. At times, these roles present different challenges. This is one of those times.

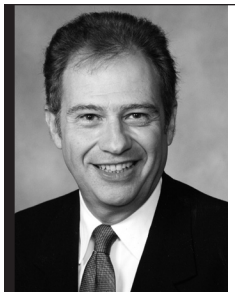
Nonetheless, as payments technology moves forward in the U.S., our payments system will continue to change as evolutionary forces generate new innovations in payments and new ways to deliver them. In some ways we will look more like the European system even as our two payments systems move to the next generation of payments. We will look more alike, although we will get there from a very different starting point. BR

The Economic Role of Cities in the 21ST Century

BY GERALD A. CARLINO

As real income increases, the demand for a greater variety of goods and services becomes a more important determinant of where people choose to live. This implies that large cities with more choices will attract high-income households that value variety. Members of these high-income households also tend to be high-skill individuals. Their presence supports cities' new function as incubators of new ideas and innovation. In "The Economic Role of Cities in the 21st Century," Jerry Carlino focuses on the economic activities that make firms in cities more productive and that make cities more attractive to urban households.

What is the role of cities in the 21st century economy? In earlier times, cities grew near transportation hubs, such as ports and railroad yards. To minimize transportation costs, firms needed to be near these hubs, and workers needed to live close to their employers to maintain reasonable commuting distances. Thus, firms and households tended to be highly concentrated in cities. These so-called agglomeration economies — the efficiency and cost savings that result from being close to suppliers, workers, and customers — were an important factor in the rise of cities as manufacturing centers.



Jerry Carlino is a senior economic advisor and economist in the Research Department of the Philadelphia Fed.

Agglomeration economies tended to support mostly the production side of the economy. That is, proximity to inputs into the production process led to gains in output. However, improvements in transportation technology mean that, today, firms are freer to locate wherever they want, and, unlike before, their choice of location will depend on where their workers choose to live. This means that an area's special features, such as its climate, will be important determinants of where households, and ultimately firms, locate.

As a result, agglomeration economies are increasingly concentrated on the consumption side. Rising real incomes mean that quality-of-life issues have become more and more important as determinants of where people choose to live. For example, growth in real income increases the demand for a greater variety of goods and services (more theaters, varied restaurant cuisine, and professional sports teams). Similarly, access to recreational amenities and better public services, such as good public schools, are also important quality-of-life issues for households.

This implies that large cities with more choices will attract high-income households that put a high value on variety. Members of these high-income households also tend to be high-skill individuals. Their presence supports cities' new function as incubators of new ideas and innovation.

To answer our question about cities' role in the 21st century economy, we will discuss some of the economic functions of cities, focusing on economic activities that make firms in cities more productive and that make cities more attractive to urban households.

AGGLOMERATION ECONOMIES IN URBAN PRODUCTION

While the discussion in this article will emphasize agglomeration economies' role in urban consumption, historically, their biggest influence has been on the production side.

Agglomeration economies constitute an important source of a firm's productivity. Increases in productivity due to agglomeration economies depend not on the size of the firm itself

(internal economies of scale),[1] but rather on the size of a firm's industry in a particular city (localization economies) or on the size of the city itself (urbanization economies).

Localization. The presence of an industry in a particular city could be the result of the available natural resources or simply historical accident. But once an industry develops in a city, other firms in that industry often reap considerable benefits by also locating there.

One advantage is sharing inputs. Consider, for example, the high-tech industry in Silicon Valley, the TV and motion picture industry in Los Angeles, and the auto industry in Detroit — three industries that have concentrated in certain locations. Many production companies in the TV industry, for example, frequently require the services of highly specialized workers, such as people who specialize in writing and editing scripts; workers who specialize in lighting, sound recording, special effects, and set design and construction; and talent agencies and firms that engage in market research.

The need to have quick access to these types of specialists is particularly important in the production of TV shows, and consequently, many of these specialists must be on or near the production set. A production company located far from Los Angeles would need to employ full-time script editors or sound and lighting personnel and set designers, for example, or else spend considerable time and money bringing them from a distance when they are needed. But when TV production companies cluster together, their combined needs for highly specialized inputs can support at least one firm that specializes in set design, others that specialize in script analysis, and so on. Thus, these services are available at lower cost from a local firm. All production companies in the cluster can enjoy a lower average cost of production by contracting for these specialized services only when they are needed.

[1] Economists have long recognized that a firm's size can affect its productivity. As a firm increases its size, it can increase productivity by having its workers specialize in particular tasks or by using its capital equipment more efficiently. In these situations, a firm is said to enjoy *internal* economies of scale.

A common labor pool allows firms to more effectively adjust their demand for labor to match fluctuations in the demand for their products.

There also are advantages to sharing a common labor pool in cities. These advantages arise from the uncertainty and variability in any one firm's demand for workers. If a firm is uncertain about the number and skill mix of workers it will hire, the firm has an incentive to cluster with other firms in its industry to draw from a common pool of workers. A common labor pool allows firms to more effectively adjust their demand for labor to match fluctuations in the demand for their products.

Consider our example of the TV industry once again. Producers of TV programs are never quite sure if a new show will be successful. But as economist Arthur O'Sullivan has noted, "When it becomes clear which programs will be discontinued, actors and technicians move from the unsuccessful programs to the successful ones. The concentration of the television industry in Los Angeles and New York facilitates the transfer of labor from one firm to another."

Common labor pools are also of value to workers as well. If any one firm in the cluster is unsuccessful and lays off workers, these unemployed workers are likely to be hired by one of the other, more successful firms in that cluster.[2]

In addition to reducing the employment risk of workers and firms, labor market pooling also facilitates the matching of workers and jobs. Having a large pool of workers in an area makes it easier for employers to find people with the set of characteristics they need. At the same time, workers are more likely to find jobs that better match their experience and skills. Therefore, having a large pool of workers in an area facilitates the number and quality of matches between firms' needs and workers' skills.

Urbanization. Not only does the size of a firm's industry in a city matter but so does the size of the city itself. Just as some kinds of businesses, such as a set-design firm, are found only where specific industries concentrate, other activities, such as financial and business services, are generally found only in urban areas. Often, only a large city can provide a client base sufficient for these specialized firms to flourish. These types of specialized services give rise to economies of scale, called urbanization economies, that are external to any one firm and its industry.

Urbanization brings greater efficiency, but it also brings problems that eventually offset the gains in efficiency. According to the traditional view, as

[2] See the article by Satyajit Chatterjee for further discussion of the advantages of labor market pooling.

cities become more congested, the increased cost of doing business (for example, in the form of higher business rents) will eventually offset any gains in agglomeration economies from additional growth. At that point, existing firms have no incentive to expand production, and new firms will not be enticed to locate in the city. The city's level of population, employment, and output will have stabilized at a certain point.

Recently, economists have focused on a new view: The creation of ideas in cities can lead to *sustained growth* in the output of urban firms even if population and employment are not expanding. The basic theory is that the higher density of population and employment in cities promotes the exchange of ideas among individuals, which economists call knowledge spillovers. The high concentration of people, especially highly skilled people, in cities creates an environment in which ideas move quickly from person to person. It's likely that some of these ideas lead to new goods and to new ways of producing existing goods.[3]

To the extent that firms more readily adopt innovations that are local, they may be able to produce more output without having to increase the level of inputs into production. In this instance, generating ideas has become an important source of growth, and proximity to individuals who create knowledge is becoming increasingly important to firms. Thus, urban locations' advantages for firms have shifted from proximity to suppliers and customers to proximity to highly skilled workers.

[3] See my 2001 *Business Review* article and my paper with Satyajit Chatterjee and Robert Hunt for further discussion of the role of knowledge spillovers in cities.

EVIDENCE ON PRODUCTION BENEFITS OF CITIES

In their 2001 research, economists Stuart Rosenthal and William Strange studied the importance of input sharing, labor market pooling, and knowledge spillovers for manufacturing firms at the state, county, and zip code levels. Among the sources Rosenthal and Strange considered, labor market pooling has a strong impact on geographic concentration of manufacturing firms at all of these levels. They also found that other types of input sharing, such as intermediate inputs and natural resources, influence the concentration of manufacturing firms at the state level but have no effect on concentration of manufacturing firms at the county or zip code levels. The effects of knowledge spillovers on the concentration of manufacturing firms tend to be more localized, influencing concentration only at the zip code level.

While Rosenthal and Strange's attempt to identify the relative importance of the various forces that gave rise to the spatial concentration of firms, the vast majority of research to date has tended to analyze the relationship between urban productivity and city size. In a 1976 study, David Segal analyzed the change in urban productivity related to the size of a metropolitan area.[4] He found that, on average, metropolitan areas with more than 2 million people are 8 percent more productive than metropolitan areas with less than 2 million people. In more productive cities, firms can afford to pay higher wages. At the same time, households and firms are drawn to relatively high productivity cities.

[4] The change in urban productivity is the amount by which output would increase as a result of increasing population in a city, with all inputs held constant.

The decline in the importance of agglomeration economies to firms does not mean that the clustering of people and jobs is no longer important to cities.

Thus, rents may also rise in these cities. In sum, if the concentration of people and jobs in cities is largely related to urban productivity, both wages and rents should increase with city size.

AGGLOMERATION ECONOMIES IN URBAN CONSUMPTION

Despite agglomeration economies' historical importance to the production side of urban economies, innovations in transportation, production, and communication technologies have weakened the economic advantage of locating closely related activities near one another. However, the decline in the importance of agglomeration economies to firms does not mean that the clustering of people and jobs is no longer important to cities. As we'll see, urban locations are still important to 21st century households.

If consumers prefer a large variety of goods and services and there are substantial economies of scale in providing them, the number of different goods and services offered and consumers' economic welfare will depend on the size of the local market.

Cultural and leisure activities offer good examples. As a hypothetical example, consider professional football, a good with relatively low per capita demand. Suppose that to break even,

the club must sell 30,000 tickets per game, or 240,000 tickets per season (based on eight home games per year). If, on average, 20 percent of a metropolitan area's residents attend a game, a metropolitan area of 1.2 million people is required to support the football team. But as a metro area's population increases, the demand for variety in professional sports teams also increases. The greater New York metropolitan area has a population of almost 20 million people and is home to nine professional sports teams in the four major sports (baseball, football, basketball, and hockey). Large metropolitan regions such as Los Angeles, Chicago, and Philadelphia support at least four teams each. With a population of only about 1 million to 1.5 million, Orlando, Hartford, and Jacksonville support one major professional sports team each (see the table).

In addition to greater variety, the quality of a good or service may improve with the population size of an area. To continue the sports analogy, economist Rodney Fort has noted that large-market teams win much more frequently than do small-market teams. The New York Yankees, a large-market team, are a post-season fixture, whereas the small-market Pittsburgh Pirates have not made the playoffs since 1992. Fort points out that teams with a large fan base earn more revenue for any given level of quality. Teams in large markets can outbid small-market teams for the best players, since large-market teams can earn more revenue from these players than do teams in small markets. The same must be true for other types of consumer goods and leisure activities, such as theaters, orchestras, and restaurants.[5]

Rising Income. In the 55 years between 1947 and 2002, per capita income adjusted for inflation (that is, real income) almost doubled in the United States. The rise in real income has led to more demand for goods and services, especially luxury goods, such as meals in gourmet restaurants and live theater, which are more plentiful in large cities.[6] Thus, the greater variety in consumption found in large cities is especially attractive to households as their wealth increases.[7] Similarly, rising incomes should increase the value that people (especially high-skill individuals) place on amenities, such as good weather.

In a 2004 study, Sanghoon Lee contended that the demand for variety may increase more than proportionately with income. That is, a 1 percent increase in income leads to more than a 1 percent increase in the demand for variety. Lee went one step further and

TABLE

Big Metro Areas Offer Diversity of Sports

Metro Area	No. of Teams	Population (Millions)
New York	9	19.9
Los Angeles	5	15.6
Chicago	5	8.6
Washington-Baltimore	5	7.2
San Francisco-Oakland-San Jose	6	6.7
Philadelphia	4	6.0
Boston	4	5.8
Detroit	4	5.4
Dallas-Ft. Worth	4	4.7
Houston	3	4.3
Atlanta	3	3.6
Cleveland	3	2.9
Pittsburgh	3	2.4
Cincinnati	2	1.9
Kansas City	2	1.7
Indianapolis	2	1.5
Orlando	1	1.5
Hartford	1	1.1
Jacksonville	1	1.0

Source: Rodney D. Fort. *Sports Economics*. New Jersey: Prentice Hall Publishers, 2003, Table 2-2. Used with permission.

[5] Leonard Nakamura discusses how innovation in retailing (introduction of scanner technology) led to larger supermarkets (superstores) that offer greater variety to their customers (bakeries, banking, pharmacies, as well as greater variety on the shelves). A number of studies by Joel Waldfogel and co-authors have shown that larger cities have more and better newspapers and more and better radio and television stations.

[6] One key feature of goods such as these is that it's difficult to transport them; therefore, they are referred to as nontraded goods and services. While people can travel to cities offering an abundance of nontraded goods and services, there is little substitution for living in the cities, or their environs, if people value convenient access to nontraded goods and services.

[7] See, for example, the articles by Jan Brueckner, Jacques-Francois Thisse, and Yves Zenou; Edward Glaeser, Jed Kolko, and Albert Saiz; and Dwight Adamson, David Clark, and Mark Partridge.

argued that since high-skill workers earn more than low-skill workers, high-skill workers will account for a larger share of the work force in large cities and a smaller share in small cities and rural areas.[8]

Other Factors. Economists Ed Glaeser, Jed Kolko, and Albert Saiz point out three other ways in which large cities enhance consumption opportunities. Large cities may provide a greater variety of public goods, too, such as more magnet schools per student (e.g., schools specializing in fine and performing arts, or those specializing in science). Furthermore, large cities make it easier for individuals to make wider social contacts and to have a more diverse set of friends. Along this line, large cities appeal to younger, more highly educated workers because large cities facilitate better development of professional and social connections than small cities and rural areas. Economists Dora Costa and Matt Kahn note that "power couples" (both partners have bachelor's degrees) are increasingly locating in large cities because large cities offer better employment opportunities for working couples. Finally, large cities may satisfy aesthetic preferences, such as the variety of architecture found in many large cities or the artistic scene in places such as New York City.

Of course, as with the production side of the urban economy, urbanization brings not only a greater variety of goods and services but also problems, such as congestion, that take the form of long-distance commuting and higher housing costs, which eventually balance the gains in variety. The higher cost of housing as cities become congested reduces households' purchasing power and limits the inflow of people.

MORE EVIDENCE ON THE BENEFITS OF CITIES

The value of a city's special traits, such as pleasant weather or the variety of consumption options, is determined by what people are willing to pay in order to live there. This amounts to the sum of what people are willing to pay for each local characteristic that adds to the quality of life in an area. The trick is to determine the prices of these local traits, since they are not bought and sold in markets.

Even though there is no explicit price for local amenities such as nice weather or greater variety, there is an *implicit* price. Suppose you are considering moving either to Metropolis, which offers its residents great variety in consumption, or to Smallville, which has far less variety than Metropolis. Because variety is something you value, you are willing to pay some extra amount, say, $1000 a year, to live in Metropolis.

You could pay your extra $1000 in two ways. One is by bidding up land prices, and ultimately rents, in Metropolis relative to Smallville. But it is not necessarily the case that you will ultimately pay $1000 more to rent a house in Metropolis. Part of the cost of living in a city with more variety could be paid in the form of wages lower than you would have accepted in Smallville. What must be true is that rent and wage differentials sum to $1000. Thus, other things equal, the extent to which rent is higher and wages are lower (so that wages adjusted for the cost of living, which economists call real wages, are lower) is the extent to which the consumption benefit of greater variety is absorbed into local land markets and local labor markets.

This discussion of a city's special traits ignores the role of the production side of the economy. Earlier we saw that if the concentration of people and jobs in cities is related to urban productivity, both wages and rents should increase with city size. But, as we just saw, if the concentration of people and jobs in cities is related to urban amenities, higher rents will outweigh higher wages, so that real wages are lower in cities offering amenities that people value.

A number of economists have looked at the relationship between a metropolitan area's size and the level of local wages and rents to determine whether productivity or urban amenities better explain the concentration of people and jobs in cities. The evidence to date is mixed. In a 2000 article, economists Takatoshi Tabuchi and Atsushi Yoshida used data for just over 100 Japanese cities for 1992 and showed that a doubling of city size is associated with about a 10 percent increase in production costs. If firms are making products for national and international markets, the only way firms in relatively high-cost (large) cities can compete with firms in relatively low-cost (small) cities is if productivity (that is, agglomeration economies) is sufficiently higher in high-cost than in low-cost cities. Thus, according to Tabuchi and Yoshida, firms in large cities incur higher costs than similar firms in small cities because large cities offer firms greater agglomeration economies.

But these authors found that a similar doubling of city size is associated with a 7 percent to 12 percent decrease in real wages, which they attribute to households' willingness to

[8] Lee's discussion ignores the role of the production side of the economy. If high-skill workers are relatively more productive than low-skill workers in cities, high-skill workers will be disproportionately drawn to large cities. Put differently, in the extreme case, highly skilled individuals may be drawn to large cities not because of the greater variety of goods and services but because such cities enhance their productivity. No doubt, both of these forces (greater productivity and greater variety) operate in cities. The difficulty is trying to differentiate the extent to which highly skilled people locate in cities because of productivity or because of greater variety.

accept lower real wages as a tradeoff for the greater variety offered in big cities. On balance, their results suggest that while productivity is higher in cities, people's taste for urban amenities and variety is an important factor in accounting for the concentration of population in cities.

In contrast, economists Gianmarco Ottaviano and Giovanni Peri studied a sample of 160 U.S. metropolitan areas and found no evidence that cultural diversity (another way to measure local variety) was important for consumers.[9] Instead, cultural diversity has a net positive impact on workers' productivity.

But the interpretation of the results of these studies assumes workers have the same level of skill to begin with; therefore, if higher real wages are found in large cities, it reflects greater productivity of similar workers in large cities. Recently, Sanghoon Lee offered another reason that real wages may differ with city size. It could be because workers with different levels of skill are attracted to different locales. For example, if real wages are found to be higher in large cities, it's not necessarily the case that agglomeration economies from locating workers together in a city are making similarly skilled workers more productive. Rather, high-skill workers, who tend to earn more than low-skill workers, may be attracted to large cities in the first place because of the higher level of amenities they offer.

As we have noted, we expect demand for variety to increase with an individual's income. Since high-skill workers also tend to earn more than low-skill workers, we expect demand for variety also to increase with a worker's skills. Given that variety increases with city size, we expect to find that high-skill workers account for a larger share of the work force in large cities and a smaller share in small cities and rural areas.

According to Lee's theory, then, it's the composition of the work force and not greater productivity that explains why wages tend to rise with city size. Lee used data from the healthcare industry to test his theory and found that large cities do, in fact, have more doctors relative to the number of nurses than do small cities. No doubt, both of these forces (greater productivity and greater variety) are at work in cities. The difficulty lies in trying to distinguish the extent to which high-wage (high-skill) workers locate in cities because large cities make them more productive or because large cities offer greater variety that high-wage workers value. This is still an open question.

Although most of the empirical results focus on the tradeoffs between wages and consumption amenities for workers, a recent study by Stuart Gabriel and Stuart Rosenthal focused on this tradeoff for firms. The researchers developed quality-of-life indexes for households and quality-of-business-environment indexes for firms in 37 cities from 1977 to 1995. They then considered how much more in wages and rents a firm is willing to pay to locate an additional worker in a city that offers the firm resources for greater productivity relative to a control city. Gabriel and Rosenthal found that many cities attractive to households are unattractive to firms (e.g., Miami, Tampa, and Albany). Similarly, they found that some cities that are attractive to firms are unattractive to households (e.g., Detroit and Washington, D.C.). Finally, a few cities were found to be attractive to both households and firms (e.g., New York, San Francisco, and Los Angeles). If the views expressed in the current article are correct, these cities are poised to do well in the new century.

CONCLUSION

Agglomeration economies will continue to play a large role in the life of 21st century cities. But unlike in earlier times, today's agglomeration economies have turned cities into centers for consumption, rather than places for manufacturing goods. In turn, this shift in focus means that cities now tend to attract more highly skilled and highly paid workers—people who want more consumption options. Consequently, modern cities must offer a wide choice of amenities to attract the high-skill workers needed in this new type of agglomeration economy.

Public policy can play a significant role in attracting and retaining highly skilled workers. Even though the productivity advantages that cities offer to firms may have waned in recent decades, the nation's largest urban areas retain many advantages in providing consumption benefits that people value. Glaeser and co-authors' 2001 study suggests that local policymakers need to focus on life-style issues because they are important in attracting and retaining high-skill workers. One such policy is providing good public schools. Other policies might focus on reducing urban crime and providing amenities such as clean streets and public parks.

[9] In their study, Ottaviano and Peri measure cultural diversity in a city as the variety of languages spoken by city residents.

REFERENCES

Adamson, Dwight W., David Clark, and Mark Partridge. "Do Agglomeration Effects and Household Amenities Have a Skill Bias?" *Journal of Regional Science*, 44 (2004), pp. 201-23.

Brueckner, Jan K., Jacques-Francois Thisse, and Yves Zenou. "Why Is Central Paris Rich and Downtown Detroit Poor? An Amenity-Based Theory," *European Economic Review*, 43 (1999), pp. 91-107.

Carlino, Gerald A. "From Centralization to Deconcentration: Economic Activity Spreads Out," Federal Reserve Bank of Philadelphia *Business Review* (May/June 1982), pp. 3-13.

Carlino, Gerald A. "Knowledge Spillovers: Cities' Role in the New Economy," Federal Reserve Bank of Philadelphia *Business Review* (Fourth Quarter 2001), pp. 17-26.

Carlino, Gerald A., Satyajit Chatterjee, and Robert Hunt. "Matching and Learning in Cities: Urban Density and the Rate of Invention," Working Paper 04-16/R, Federal Reserve Bank of Philadelphia (2004).

Chatterjee, Satyajit. "Agglomeration Economies: The Spark That Ignites a City?" Federal Reserve Bank of Philadelphia *Business Review* (Fourth Quarter 2003), pp. 6-13.

Clement, Douglas. "Urban Legends," Federal Reserve Bank of Minneapolis: *The Region*, 18 (2004).

Costa, Dora L., and Matthew E. Kahn. "Power Couples," *Quarterly Journal of Economics*, 115 (2000), pp. 1287-1315.

Fort, Rodney D. *Sports Economics*. Upper Saddle River, New Jersey: Prentice Hall, 2003.

Gabriel, Stuart A., and Stuart S. Rosenthal. "Quality of the Business Environment Versus Quality of Life: Do Firms and Households Like the Same Cities?" *Review of Economics and Statistics*, 86 (2004), pp. 438-44.

Glaeser, Edward L., and Albert Saiz, "The Rise of the Skilled City," Brookings-Wharton Papers on Urban Affairs 5 (2004), pp. 47-94.

Glaeser, Edward L., Jed Kolko, and Albert Saiz. "Consumer City," *Journal of Economic Geography*, 1 (2001), pp. 27-50.

Lee, Sanghoon. "Ability Sorting and Consumer City," unpublished manuscript, University of Minnesota and Federal Reserve Bank of Minneapolis, 2004.

Moretti, Enrico. "Estimating the Social Return to Higher Education: Evidence from Longitudinal and Repeated Cross-Sectional Data," *Journal of Econometrics*, 121(2004), pp. 175-212.

Nakamura, Leonard. "Is the U.S. Economy Really Growing too Slowly?" Federal Reserve Bank of Philadelphia *Business Review* (March/April 1997), pp. 3-14.

Ottaviano, Gianmarco, and Giovanni Peri. "The Economic Value of Cultural Diversity: Evidence from U.S. Cities," National Bureau of Economic Research Working Paper 10904 (November 2004).

O'Sullivan, Arthur. *Urban Economics*. Boston, MA: McGraw-Hill Irwin, 2003.

Rauch, James E. "Productivity Gains from Geographic Concentration of Human Capital: Evidence from Cities," *Journal of Urban Economics*, 34 (1993), pp. 380-400.

Rosenthal, Stuart S., and William C. Strange. "The Determinants of Agglomeration," *Journal of Urban Economics*, 50 (2001), pp. 91-229.

Segal, David. "Are There Returns to Scale in City Size?" *Review of Economics and Statistics*, 58 (1976), pp. 339-350.

Tabuchi, Takatoshi, and Atsushi Yoshida. "Separating Urban Agglomeration Economies in Consumption and Production," *Journal of Urban Economics*, 48 (2000), pp.70-84.

Waldfogel, Joel. "Who Benefits Whom in Local Television Markets?" Brookings-Wharton Papers on Urban Affairs, 2003.

Waldfogel, Joel, and Lisa George. "Who Affects Whom in Daily Newspaper Markets?" *Journal of Political Economy*, 2003.

Waldfogel, Joel, and Peter Siegelman. "Race and Radio: Preference Externalities, Minority Ownership, and the Provision of Programming to Minorities," *Advances in Applied Microeconomics*, 10, 2001.

The Economics of Asset Securitization

BY RONEL ELUL

Asset securitization — transforming illiquid assets into tradable securities — is a large and growing market, even rivaling the corporate debt market in size. While the underlying assets can be very different — ranging from song royalties to home mortgages — most asset-backed securities nevertheless share some distinctive features. In "The Economics of Asset Securitization," Ronel Elul explains why asset-backed securities exist and discusses some reasons for their common structure.

In 1997 rock star David Bowie raised $55 million by selling bonds backed by revenues from his first 25 albums.[1] This was the first application of securitization to intellectual property. Formally speaking, *asset securitization* refers to the process whereby nontraded assets — such as song royalties — are transformed into tradable securities, called asset-backed securities, or ABS, through the repackaging of their cash flows. Some more mainstream examples of asset-backed securities include mortgage-backed securities (MBS) and secured credit card receivables.

Securitization is a large and growing market. Currently, it represents about 25 percent of new nongovernment borrowing.[2] To take just one of the sectors mentioned above, at the end of 2003 there was more than $7 trillion in securitized mortgages, representing nearly three-quarters of all outstanding home loans.

While the underlying assets can be very different (in terms of maturity, collateral, and risk, for example), ABS nevertheless tend to share some common features. These common elements, which we discuss in further detail below, include selling the underlying assets so that they are moved off the firm's balance sheet, grouping individually illiquid assets into portfolios, taking steps to reduce the risk of default on the underlying assets (known as credit enhancement), and subdividing the assets into several classes of securities (tranching). Financial economists have attempted to explain the underlying reasons for securitization, as well as these common features.

MORTGAGE-BACKED SECURITIES: AN EXAMPLE OF ASSET SECURITIZATION

Consider, for example, a bank (the originator) that offers a $200,000 mortgage to a home buyer (see Figure) with an interest rate of 6 percent. Rather than hold this loan in its portfolio and receive small monthly payments for a period of 30 years, the bank may prefer to move the loan off its balance sheet by selling it to an outside investor. In this way the bank receives funds today from selling the loan, so that it has the opportunity to profit further by originating even more loans; the reason is that the bank typically collects a fee (the origination fee) for each loan it originates.[3] There are also other motivations for securitization that we will discuss below. But for now, let's look at how the bank in our example might use securitization.

[1] Despite initial predictions, this has not led to a wave of such issues, in part because the Bowie bonds themselves have not performed quite as well as expected (because online music piracy has curtailed revenues from music sales).

[2] Further detail can be found in the Flow of Funds Accounts tabulated by the Federal Reserve Board.

[3] A typical fee is 1 percent of the loan amount.

Ronel Elul is a senior economist in the Research Department of the Philadelphia Fed.

The problem is that an individual loan is very illiquid, i.e., hard to sell, in part because potential buyers know much less about the homeowner than does the bank. For example, the bank probably knows more about its own underwriting standards than any potential buyer, or the bank may have had a prior lending relationship with the borrower. Instead of selling the entire loan to an individual buyer, the bank can agree to sell all or most of its loans to an issuer — typically a government sponsored enterprise (GSE) such as Fannie Mae or Freddie Mac — that pools these loans with ones made by other lenders (see Figure). For example, rather than a single $200,000 mortgage, the pool may consist of $600,000 in mortgages — that is, three such loans.[4] This means that instead of buying 100 percent of a single mortgage, a potential investor who has $200,000 to spend may end up with a claim on one-third of each mortgage.

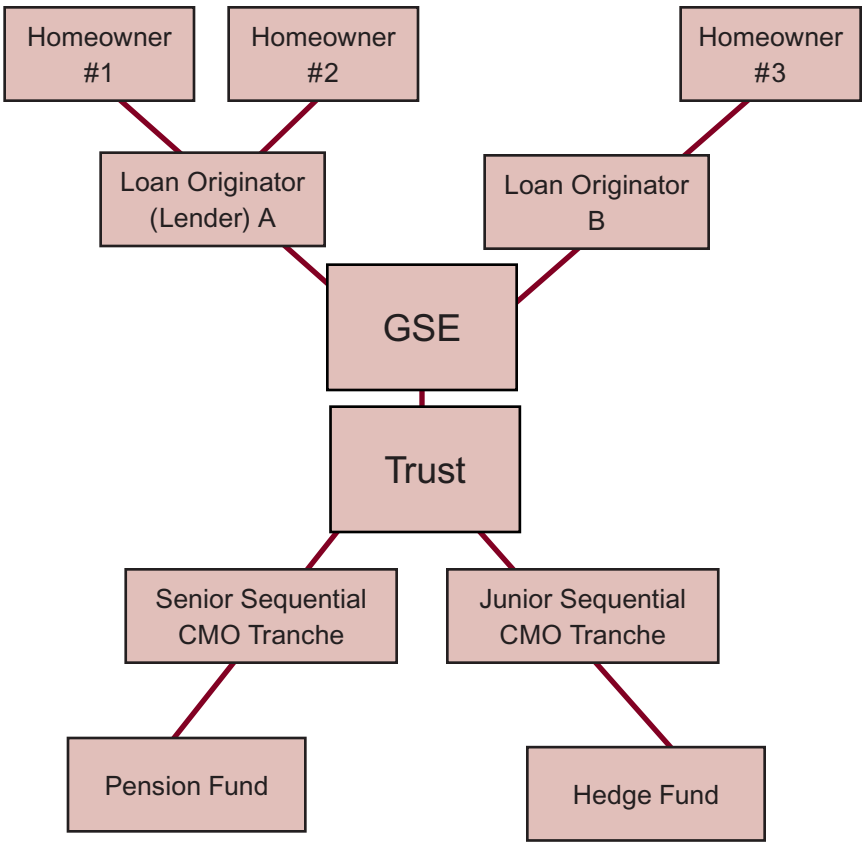
The GSE will place these mortgages in a trust (also known as a *special-purpose vehicle*) (see Figure) and then insure the pool against default; this is a form of *credit enhancement*, a technique for improving the credit quality of one or more of the vehicle's assets. Credit enhancement can take several forms: overcollateralization (so that the dollar value of the assets in the pool exceeds the value of the securities issued), the use of a GSE or other outside insurer to guarantee payment, and tranching, which we discuss later. In many securitizations more than one of these may be used.

The trust then issues securities, known as *mortgage-backed securities* (MBS), against this pool. Like other bonds, these securities promise the buyer regular interest payments and the return of principal at maturity, and they are financed from the cash flows of the underlying mortgages. Notice that when the assets are moved off balance sheet, they are legally separated from the bank that originated the mortgages, so that creditors of the bank (such as depositors and its bondholders) do not have any claims on these assets, and investors who receive mortgage payments do not have any claims on the originating bank. A certain amount is deducted from the monthly payments on the mortgages before they are passed through to the investor; this money covers the servicing of the mortgages (i.e., collecting the monthly payments, which is often done by the issuing bank) and also serves as compensation to the GSE for its guarantee. For instance, in our example, although homeowners pay an interest rate of 6 percent, investors may receive only 5.5 percent.

Investors will usually find it more attractive to purchase an MBS than to purchase an individual mortgage loan. First, investors are exposed to much less risk because the pooling process diversifies away the impact of an individual mortgage's performance. For example, investors do not need to worry as much about an individual homeowner's behavior (although economy-wide disturbances that affect many homeowners at once will still be important). Second, the securities are

FIGURE

The Securitization of Mortgages



[4] In practice, a typical pool may consist of several hundred loans and have a face value of $50 million.

also much more liquid than individual mortgages because the pooling process makes each MBS much more similar to its peers; that is, pooling makes the characteristics of an individual loan much less important to potential investors. This reduces the amount of information potential investors need to collect before purchasing the security and thereby makes it easier to trade.

Finally, the issuer of the MBS may also further manipulate the cash flows from the pool of mortgages by splitting them into classes known as *tranches* (see Figure).[5] The difference between one tranche and another varies depending on the type of asset securitized. In the case of mortgage-backed securities, tranches are often structured in terms of principal payments on the mortgages in the pool. That is, the structure is used to allocate *prepayment risk*, the risk that a security will pay off before its maturity date, thereby forcing the investor to reinvest his funds at a (possibly) lower rate. The simplest structure is known as "sequential pay" (more complex ones are also used). As the name suggests, in this case the tranches are retired in sequential order. That is, investors in the first — *senior* — tranche receive principal payments from the underlying assets first, those in the second tranche next, and so on. Investors in the last — most junior — *tranche* receive principal payments from the mortgages in the pool only when the tranches ahead of them in priority have been fully paid.

For instance, suppose that in our example, the $600,000 pool consisting of three mortgages was divided into two tranches: a senior one with a principal balance of $200,000 and a more junior one with a balance of $400,000. Then if all mortgages paid according to schedule, it would take 16.5 years for the senior tranche to be paid down.[6] During this time, the senior tranche would receive all of the principal payments on the mortgages in the pool, as well as interest payments of 5.5 percent on its outstanding balance. The junior tranche would receive only its interest payments. After the senior tranche has been fully paid down, the junior tranche would then begin to receive principal payments and would be fully retired after 30 years.[7]

Now suppose that shortly after the mortgages are issued, one of the homeowners sells his house and pays off his mortgage. In this case, the senior tranche is paid off immediately. The junior tranche would then begin to receive principal payments as well; nevertheless, so long as the other mortgages do not pre-pay, it would still take 30 years to fully pay down this tranche. Notice that the junior tranche is thus much less sensitive to prepayment risk than the senior tranche.

However, in other ABS, the absence of a GSE guarantee means that the determining factor in structuring the tranches is typically credit risk; that is, a senior tranche would have priority over a junior one in the event of a default, so that it has first claim on the securitization's underlying assets. As a result, tranching can serve as a form of credit enhancement; in particular, it enhances the credit quality of the more senior tranches at the expense of the junior ones (the senior tranche is typically AAA-rated in these cases).[8]

In this example we can see the key features of asset securitization: a sale of the underlying assets so that they are moved off the issuer's balance sheet, the pooling of illiquid assets, credit enhancement, and tranching.

The issuer of the MBS may also further manipulate the cash flows from the pool of mortgages by splitting them into classes known as tranches.

WHAT ASYMMETRIC INFORMATION CAN TELL US ABOUT ASSET SECURITIZATION

When Investors Are Uninformed, Capital Structure Matters. A firm's decision about whether — and how — to securitize assets can be viewed as a variant of the broader question of how a firm should finance

[5] In the case of mortgages, a tranched security is known as a REMIC (real estate mortgage investment conduit) or CMO (collateralized mortgage obligation).

[6] This figure can easily be obtained from any online mortgage amortization calculator; several such calculators are available.

[7] Notice that another implication of the sequential pay structure is that the senior tranche has a shorter maturity than the underlying mortgages; thus, tranching also facilitates participation in this market by investors with shorter investment horizons.

[8] Bonds are rated according to their default risk by ratings agencies, the most prominent of which are Moody's and Standard & Poor's. Although each agency uses slightly different classifications, ratings are assigned in alphabetical order, with AAA being the least risky (Aaa for Moody's) and D representing a bond that is in default. Bonds rated BBB or above by Standard & Poor's (Baa for Moody's) are termed "investment grade."

itself. This is known as the capital structure decision.

In 1958, future Nobel Prize winners Franco Modigliani and Merton Miller showed that the form of financing a firm uses does not affect the total value of its assets under a number of particular assumptions. This is known as the Modigliani-Miller proposition. Some key assumptions — which we will revisit below — are that corporate bankruptcy is costless, that there are no applicable government regulations, and that all types of securities have similar tax treatment. Another important assumption is that outside investors are as well informed as the firm's insiders (such as management) about the firm's prospects. When this is true, insiders and outsiders are said to be *symmetrically informed*.

On the other hand, when insiders know more than outside investors (which is often a more realistic assumption), the mix of debt securities[9] and equity (that is, stock) — and who holds each — can affect the firm's ability to secure funds from outside investors and, ultimately, the value of the firm itself. Two classic papers examine these issues, and the ideas in these articles can also be used to explain some of the key features of ABS.

In their article, economists Hayne Leland and David Pyle explain why insiders tend to retain an equity stake in their firm, rather than selling all of the firm's shares to the public. Insiders who believe that a firm's future profits are likely to be high would like to convince skeptical investors. On the other hand, skeptical investors believe that talk is cheap. They reason that insiders are simply trying to sell stock in the firm at the highest possible price, whatever the firm's true prospects. However, insiders can credibly signal their information to the market by holding a larger share of the firm's stock. In effect, an insider who holds a significant ownership stake is putting his money where his mouth is. This allows the firm to sell its stock at a higher price but will leave insiders exposed to more risk because their ownership share in the firm keeps them from holding a well-diversified portfolio; this increased risk is the cost insiders must bear to gain credibility.

The problem firms face when issuing equity is that outsiders are understandably suspicious that insiders know something they do not and that the stock is overvalued.

An article by economists Stewart Myers and Nicholas Majluf explains why firms often prefer to sell debt securities rather than issue equity to outside investors. The problem firms face when issuing equity is that outsiders are understandably suspicious that insiders know something they do not and that the stock is overvalued. As a result, the firm can increase the price investors are willing to pay for its securities by offering securities that are *informationally insensitive*, that is, securities whose payoffs do not depend on factors known only to insiders.

For example, since debt payments are contractually fixed whether the firm's profit is high or low, debt is less informationally sensitive than equity; therefore, the firm can secure outside funds at a lower cost by issuing bonds rather than stock.[10]

I'll now show how these ideas can help to explain some of the distinctive features of securitization.

Tranching Allows Issuers to Sell Safe Cash Flows and Retain Risky Ones. Suppose an issuer (for example, a bank) has a portfolio of assets such as credit card receivables, that is, expected payments on credit card balances. This portfolio is not as liquid as the issuer would like, and so the issuer might prefer to sell part of it for cash through a securitization. However, the issuer's information about the quality of its assets is superior to that of potential investors, perhaps because the bank has proprietary information about its customers that it has collected over a long period. Having such information makes any sale costly and difficult. The bank's goal is to structure the security so as to maximize its revenue from selling the assets.

Economists Peter DeMarzo and Darrell Duffie show that to maximize revenue, the issuer should sell a senior tranche backed by the assets while retaining the junior tranche. By analogy with the firm's capital structure

[9] Corporate and government bonds are common examples of debt securities. A debt security represents the issuer's promise to repay the loan's face amount, with interest, in a set period of time. By contrast, the firm is under no contractual obligation to pay shareholders dividends of any set amount.

[10] This is known as the Myers-Majluf pecking order theory because the firm has a "pecking order" of financing choices. It relies as much as possible on retained earnings (which bypasses outside investors completely). If retained earnings do not suffice to finance its projects, it issues debt. Only if the firm does not have the earnings to make debt payments does it issue equity to outside investors (a start-up firm might fall into this category).

decision, the most junior tranche is also often termed the *equity* stake. Moreover, they show that the higher the quality of the assets, the larger this retained *equity* stake. This follows the work of Leland and Pyle in that the issuer signals that its assets are of high quality by holding an equity stake; it is also reminiscent of Myers and Majluf's model in that an informationally insensitive security is issued to uninformed outside investors.

To take a recent example, which is fairly typical, in a 2002 credit card securitization by Fleet Bank (now part of Bank of America), the issuer retained an equity interest equal to approximately 10 percent of the total principal.

Peter DeMarzo further extends this model to explain why we often see pooling of assets (recall that this is a distinctive feature of many securitizations) before tranching occurs. DeMarzo shows that pooling assets involves a tradeoff. On the one hand, by selling different assets as a single unit, the issuer cannot signal information about the asset by retaining a specific amount of equity for each individual asset. On the other hand, to the extent that pooling diversifies idiosyncratic risk, it allows the issuer to sell a larger quantity of informationally insensitive securities.[11] When the benefits from diversification outweigh the limitations of selling the assets together (for example, when the issuer has many similar mortgages available), then pooling is beneficial.

Tranching Increases Information Production by Investors. While DeMarzo and Duffie's model provides useful insights, its underlying assumptions do not reflect significant parts of the ABS market. In many cases, investors may actually know at least as much about the assets as the issuer, and even more significantly, some potential investors may know more than others (this is the case for mortgage-backed securities, for example). Of course, investors do not receive this information for free. Hedge funds that specialize in buying mortgage-backed securities must pay substantial salaries to Ph.D.s who understand these securities.

In many cases, investors may actually know at least as much about the assets as the issuer, and even more significantly, some potential investors may know more than others.

Economists Arnoud Boot and Anjan Thakor develop a model in which sellers of ABS exploit the fact that potential investors may choose to invest in learning about the underlying assets. Boot and Thakor show that both the pooling and tranching of assets can encourage investors to learn about these assets, so that they are willing to pay more for them.

Their idea is that by separating the cash flows from the asset into senior and junior tranches, the issuer creates a highly informationally sensitive security — the junior tranche. Since a junior tranche is riskier, investors need to learn more about the assets underlying this junior security in order to determine whether it is worth buying. By contrast, a high-rated senior tranche carries less risk, so that even uninformed investors can safely invest in it.

This structure maximizes incentives for sophisticated investors to become informed about the value of the underlying assets, since such investors can specialize in buying only this most informationally sensitive portion of the cash flows. Conversely, uninformed investors purchase the informationally insensitive senior tranche. Also note that unlike in DeMarzo and Duffie's model, the issuing firm itself does not need to retain anything, since it knows nothing more than investors do.

Boot and Thakor also offer a similar explanation for why securitizations often involve the pooling of assets. The reason is that the risks of the assets pooled in the ABS have two components: a common one (such as interest-rate risk or national price trends in the case of mortgages) and an idiosyncratic one (e.g., a particular borrower's individual default risk). Pooling assets makes acquiring information more effective because the idiosyncratic risk is diversified and investors can concentrate their efforts on learning about the common characteristics of these assets without worrying that their efforts will be undone by an individual homeowner's unpredictable finances.

Economist Guillaume Plantin provides evidence that in collateralized debt obligations,[12] it indeed appears as if sophisticated investors, such as hedge funds, purchase the more junior "equity" tranches, whereas relatively unsophisticated investors specialize in

[11] Idiosyncratic risk is risk related to the unique circumstances of a specific loan or borrower, as opposed to overall market risk, which affects many assets at once.

[12] Collateralized debt obligations, or CDOs, are securities in which the underlying assets are themselves loans or bonds, most typically risky corporate debt ("junk bonds").

the high-rated senior tranches, commonly known as "A" tranches.[13]

Structures with Many Tranches. In the models discussed above, the resulting structure of the securitization is very simple: usually only two tranches, one senior and one junior. In practice, most structures are somewhat more complicated and feature multiple tranches. For example, in the Fleet credit card securitization discussed earlier, there were actually *three* tranches: a senior AAA-rated "A" tranche, a more junior "B" tranche (which was A rated), and the unrated equity tranche. Plantin's paper explains why these multiple tranches might arise; he also demonstrates that — as in the papers by Boot and Thakor and DeMarzo and Duffie — the optimal structure is a senior-junior securitization in which the higher-rated senior tranches have absolute priority over the low-rated junior ones in the event of a default.

Plantin's model features multiple tranches because it includes several classes of potential investors with different degrees of sophistication (for example, hedge funds, pension funds, and individual investors). For Plantin, a sophisticated investor is more likely to discover when a given pool of assets is worth buying, whereas a less sophisticated investor is more likely to remain uninformed. Having multiple investors that differ in their sophistication allows for multiple tranches in the optimal structure.

Plantin produces useful insights by explicitly modeling the sale of ABS as an auction. Auctions are the common sales method when securities are privately placed (as opposed to being publicly issued).[14] The auction may be informal, in which case the issuer privately consults each potential buyer before choosing the best offer. Alternatively, if there are many potential bidders, a formal auction may be used, typically a first-price sealed-bid auction.[15] In either case, economists have a well-developed set of insights about the forces at play in an auction.[16]

In particular, in an auction where each bidder has his own information about the true value of the items being sold, there is the risk that the buyer who wins the auction is the one who has overpaid. This is known as the *winner's curse*. This problem should be familiar to anyone who has won an eBay bidding war, only to later discover that the item is available for retail purchase at a lower price. Note that the winner's curse is not the result of bidders' allowing their emotions to get the better of their reason. Rather, it arises because bidders are not equally well informed about the valuation of the object (in this example, the price for which it can be bought elsewhere). A rational bidder takes this into account when bidding. As a result, instead of bidding his estimate of the object's value, the bidder will shave down his bid to reflect the fact that he is likelier to win when he has overestimated the value of the object.

In an auction where each bidder has his own information about the true value of the items being sold, there is the risk that the buyer who wins the auction is the one who has overpaid.

In Plantin's model, the issuer would like to maximize participation in the auctions for the securities he offers. The reason is that the more potential bidders there are, the likelier it is that some bidder will receive information that confirms that these assets are indeed of high quality, in which case he would be willing to pay a high price. In particular, the issuer would like to encourage sophisticated investors to participate, since they are likeliest to receive information concerning the asset. On the other hand, the more sophisticated investors there are, the more severe the winner's curse. The reason is that those investors who are not informed know they will win the auction only if none of the other investors learn that the assets are of high quality. If many of the other investors are sophisticated, the absence

[13] For example, banks are among the most active buyers of higher-rated senior tranches. The reader may find it strange to think of banks as unsophisticated, but Jianping Mei and Anthony Saunders have demonstrated that — at least in the case of real estate loans — banks seem to act naively in lending on the basis of past returns rather than expected future performance. Of course, degrees of sophistication need not explain why banks favor the senior tranches — there are regulatory reasons for banks to invest in less risky securities.

[14] In a private placement, securities are issued to "qualified institutional investors" (such as insurance companies), rather than to the general public, as in a public offering. The advantage is that there is much less regulation; the disadvantage is that since there is a very limited secondary market, the price received is typically lower. The "Bowie bonds" discussed earlier were privately placed; Prudential Insurance Company purchased the entire issue. More generally, private placements make up approximately 15 percent of all nonmortgage ABS issued.

[15] In a first-price sealed-bid auction, each bidder submits a sealed bid to the seller (a bid that is hidden from other bidders). The high bidder wins and pays his bid for the good. Generally, a sealed-bid format has two distinct parts: a bidding period in which participants submit their bids, and a resolution phase in which the bids are opened and the winner determined.

[16] See, for example, the book by Paul Klemperer and the book by Paul Milgrom.

of higher bids suggests that the assets are indeed of very low quality. Thus, the uninformed investors are timid in their bidding, which will reduce the issuer's revenue from the auction. Thus, designing the structure so as to encourage more sophisticated investors to participate in the auction creates a tradeoff: Sophisticated investors — who are likelier to be well-informed about the assets — will bid more aggressively and so will pay a higher price for high-quality assets. But they exacerbate the severity of the winner's curse for the uninformed investors and make them timid bidders.

Tranching plays a dual role in resolving this tradeoff. It draws in sophisticated investors by creating an informationally sensitive junior tranche, as in Boot and Thakor's model. Since Plantin assumes that sophisticated investors must bear a higher cost to participate in the auction for any given tranche, these investors focus their efforts only on the most junior tranche.[17] By contrast, unsophisticated investors participate in the auctions for all of the tranches. Since the sophisticated investors bid for only the most junior tranche, the unsophisticated investors can bid aggressively for the senior tranches without fear of the winner's curse, which increases the issuer's revenue.[18] While these unsophisticated investors also bid for the most junior tranches, the winner's curse means that they do so very conservatively and therefore are less likely to end up holding these tranches when the auction closes. This is consistent with the empirical evidence presented earlier: junior tranches do indeed seem to be held by more sophisticated investors.

REGULATION: ANOTHER DRIVER OF SECURITIZATION

Legal factors and government regulation are also important drivers of securitization. Three main regulatory and legal forces encourage securitization and determine some of its characteristics.

When a bank securitizes mortgages, investors in the mortgage-backed securities are virtually guaranteed that they will be paid in full, regardless of how the bank itself fares in the future.

Securitization May Reduce Bankruptcy Costs. As mentioned above, securitization is typically off balance sheet in that the underlying assets are legally separated from the firm so that the firm's creditors do not have any claim on these assets. Recall that the Modigliani-Miller proposition assumed that bankruptcy is costless. In practice, of course, it is not. Bankruptcy costs take two forms: direct costs, such as lawyers' fees and court costs, and indirect costs, which include difficulties in raising funds to make profitable investments, inefficient investments undertaken while in bankruptcy, and so on. These indirect costs may also affect a firm when it is in financial distress, that is, even when it is close to bankruptcy. Investors (both shareholders and creditors) will, of course, ultimately bear these costs because the value of their securities will be impaired in bankruptcy. Anticipating these costs, investors will be more reluctant to offer funds in the first place, which will raise the firm's cost of financing (since they will obtain a lower price for any securities they offer).

Economists Gary Gorton and Nicholas Souleles point out that moving assets off balance sheet can be helpful because firms can mitigate these bankruptcy costs by precluding creditors' access to these assets.[19] For example, when a bank securitizes mortgages, investors in the mortgage-backed securities are virtually guaranteed that they will be paid in full, regardless of how the bank itself fares in the future. Consequently, they are willing to offer a high price for these securities. By contrast, if the bank retains the mortgages, investors will share in both the cash flows from the assets and the costs the issuer incurs should it find itself in financial distress. As a result, investors offer a relatively lower price for these securities. This is particularly true for risky, low-rated issuers. A classic example is Chrysler: It successfully used securitization in a period of financial distress (1990-91) when it could neither finance car loans in the commercial paper market nor issue long-term debt.[20]

However, not every type of asset lends itself to securitization. Economists Kenneth Ayotte and Stav Gaon

[17] Plantin argues that this is because it is difficult for sophisticated investors to find retail clients to ultimately hold these securities; for example, only wealthy "qualified investors" are permitted to invest in hedge funds.

[18] The idea that creating a riskless security can encourage participation by uninformed investors was first used by Gary Gorton and George Pennacchi to explain how insuring bank deposits protects uninformed investors and thereby makes them willing to fund banks.

[19] Bankruptcy costs are also further reduced by the credit enhancement that is a feature of nearly all securitizations.

[20] See the article by Dennis Cantwell.

show that if the assets are essential for the firm's continuing operations, the firm's losing control over them through a securitization may imperil the firm's existence in case of financial distress. The reason is that the holders of the securitized assets have little interest in the firm's continued survival and may not be willing to compromise to help the firm avoid liquidation. Ayotte and Gaon offer the example of the bankrupt steel firm LTV, which made this argument as part of an attempt to regain control of inventory it had securitized.[21]

Securitization Can Lower Banks' Regulatory Capital Requirements. Some economists have argued that bank capital requirements are important drivers of securitization. This is also known as *regulatory arbitrage* because securitization might allow banks to shift assets to lower their minimum regulatory capital requirements. In particular, to the extent that minimum capital requirements do not assign to each asset the capital that would be held by an unregulated financial intermediary, it might be profitable for banks to sell off low-risk loans (such as mortgages) and retain high-risk assets. Note that for this to be an effective "arbitrage," the loan's buyer must have a lower capital requirement for holding that loan than the selling bank (for example, an unregulated hedge fund). As long as this is true, it is cheaper for the buyer to hold the loan on its books than for the bank, and both can profit from its sale.[22]

Consider the following example. A bank can make one of two $100,000 loans, both of which require 8 cents of capital per dollar lent.[23] One loan is an adjustable-rate mortgage with an 80 percent loan-to-value ratio. In 2000 the interest rate on such a mortgage averaged 7 percent, and the default rate was approximately 0.5 percent. The other loan is a small-business line of credit, with an interest rate of 7.4 percent and a default rate of 1.5 percent.[24] Notice that the expected return on the mortgage can never be higher than 7 percent. By contrast, the small-business loan has an expected return that is at least 7.29 percent.[25] Given the regulatory capital requirements, the bank may prefer to hold the risky small-business loan and sell the safe mortgage. The reason is that under current minimum capital requirements, both loans require the bank to hold $8,000 of capital, but the high-risk small-business loan has an expected return that is nearly 30 basis points higher.

The evidence as to whether regulatory arbitrage is an empirically significant driver of securitization is mixed. On the one hand, Brent Ambrose, Michael LaCour-Little, and Anthony Sanders do find evidence consistent with regulatory arbitrage in the mortgage market. By contrast, however, Bernadette Minton, Anthony Sanders, and Philip Strahan provide empirical evidence that casts doubt on the importance of regulatory arbitrage and instead supports the hypothesis that securitization is motivated by a desire to reduce bankruptcy costs. In particular, they find that unregulated issuers (which are not subject to capital requirements) seem to be more active securitizers than banks. Moreover, it is the riskier firms (for which bankruptcy is obviously more of a concern) that use securitization the most.

Pension Fund Regulations Can Explain Credit Enhancement and Tranching. Finally, regulations governing pension funds are also important for securitization. The most prominent of these regulations are found in the Employee Retirement Income Security Act (ERISA). ERISA regulations govern pension funds' investment portfolios. Among these regulations are those that restrict funds' holdings of low-rated or very junior asset-backed securities in certain circumstances. This clearly encourages the use of credit enhancement in ABS structures and, in particular, the creation of high-rated senior tranches.[26] In light of the regulations' obvious importance, it is somewhat surprising that economists have yet to examine their relative weight in the growth of securitization.

IS SECURITIZATION EFFICIENT?

One important question we have not yet discussed is the social implications of securitization. That is, does it provide a net benefit to society or perhaps simply lead to a transfer of wealth from one party to another? Said differently, many of the models

[21] Gorton and Souleles suggest that another reason firms may not want to securitize all assets is that interest payments on off-balance-sheet debt are not always tax-deductible to the issuing firm (although in practice lawyers have developed structures that allow the tax advantages to flow back to the issuer).

[22] In addition to minimum capital requirements, bank regulators can also limit regulatory arbitrage through the examination process.

[23] In rough terms, the capital requirement means that for each dollar lent, the bank must secure at least 8 cents of funding in the form of retained earnings, stock, or long-term subordinated debt (i.e., debt that is junior to deposits).

[24] The data on small-business loans are from the paper by Sumit Agarwal, Souphala Chomsisengphet, and Chunlin Liu.

[25] Because the small-business loan repays 100-1.5 percent = 98.5 percent of the time. As a result, even if this loan returns nothing when it defaults, its expected return is at least 0.985 × 7.4 percent = 7.29 percent.

[26] Many institutional investors also have self-imposed restrictions on the credit quality of their portfolio.

we have discussed involve the issuer's structuring the securitization so as to maximize his revenues. But is the issuer's gain merely the investor's loss?

Securitization Can Be Socially Beneficial. Recall that bankruptcy costs seem to be an important driver of securitization (explaining its off-balance-sheet feature as well as the credit enhancement). This ability to mitigate bankruptcy costs is certainly likely to be beneficial; we have already seen, for example, that securitization helped Chrysler Corporation continue operating during a time of financial distress.

In many of the other models we examined, securitization is also implicitly beneficial, since it is structured so as to reduce information asymmetries. That is, investors may be willing to pay more for certain tranches either because they are more confident that the securities they are buying are of high quality or because the structure makes it more profitable for them to become informed about the assets. In either case, this lowers the cost of financing for the firm and could allow it to fund profitable projects that might otherwise be infeasible. This is good for society; everyone can be made better off if profitable projects are not forgone.

Securitization May Sometimes Be Harmful. Having said this, however, securitization could potentially have social costs for several reasons.

To the extent that securitization permits firms to circumvent bankruptcy law or to circumvent banks' minimum capital requirements, it is unlikely to be socially optimal. In addition, the recent example of Enron has shown that securitization can sometimes be used to facilitate fraud. By moving assets and liabilities off its balance sheet, Enron was able to muddy investors' picture of the firm. Enron also implicitly guaranteed some of the assets it securitized, so that they were not truly off balance sheet. As a result, the firm was actually much riskier than it appeared.

Finally, in some of our models, securities were structured so as to maximize the investors' incentives to become informed. While this may sometimes facilitate the funding of projects that otherwise would not be financed, it provides no net social benefit if the project would have been financed without securitization. Moreover, by driving potential buyers to spend money on acquiring information, the issuer would actually be encouraging unnecessary investment in information production.[27] To put it another way, society as a whole would be better off if the assets were simply sold without being subdivided into tranches, and as a result, investors did not need to invest the resources necessary to purchase these junior tranches.

CONCLUSION

Securitization is a large and growing area of corporate finance. Its key features are that it is typically off balance sheet, combines many small assets into a pool, and often divides this pool of cash flows into tranches.

According to some theories, off-balance-sheet financing and, to a limited extent, tranching are responses to government regulations. Bankruptcy costs also help explain why securitization is conducted off balance sheet and also why it commonly features credit enhancement.

Another set of explanations we have explored is based on the existence of differences in information about the underlying assets — either between issuers and potential investors or between different classes of investors. These theories show that securities may be designed to alleviate these differences in information, so that outside investors are comfortable purchasing them, and they may also be designed to encourage investors to become better informed about the underlying assets. This is manifested in the pooling of assets and the subsequent division of these cash flows into tranches.[28]

According to some theories, off-balance-sheet financing and, to a limited extent, tranching are responses to government regulations.

While there is a well-developed body of theoretical work that explores the determinants and structure of securitization, the empirical significance of these models, and in particular the impact of government regulation and bankruptcy law on securitization, remains a ripe area for future research.

[27] This is similar to the argument often made against advertising.

[28] Information asymmetries and regulations are not the only explanations for why new securities are introduced. There is an interesting literature in which securities are designed to fill unmet needs for risk-sharing, that is, to *complete markets*. For example, a futures contract allows farmers to lock in a price for wheat so that they are not exposed to the risk that prices will collapse. For a model in which completing markets drives financial innovation, see Franklin Allen and Douglas Gale.

REFERENCES

Agarwal, Sumit, Souphala Chomsisengphet, and Chunlin Liu. "Determinants of Small Business Default," Working Paper, University of Nevada, Reno, 2004.

Allen, Franklin, and Douglas Gale. "Optimal Security Design," *Review of Financial Studies*, 1, 1988, pp. 229-63.

Ambrose, Brent, Michael LaCour-Little, and Anthony Sanders. "Does Regulatory Capital Arbitrage or Asymmetric Information Drive Securitization?" Working Paper, Ohio State University, November 2003.

Ayotte, Kenneth, and Stav Gaon. "Asset-Backed Securities: Costs and Benefits of 'Bankruptcy Remoteness'," manuscript, 2004.

Boot, Arnoud, and Anjan Thakor. "Security Design," *Journal of Finance*, 48, 1993, pp. 1349-78.

Cantwell, Dennis. "How Public Corporations Use Securitization in Meeting Financial Needs: The Case of Chrysler Corporation," in Kendall, Leon T., and Michael J. Fishman (eds.): *A Primer on Securitization*. Cambridge, MA: MIT Press, 1996.

DeMarzo, Peter, and Darrell Duffie. "A Liquidity-Based Model of Security Design," *Econometrica*, 1999, 67, pp. 65-99.

DeMarzo, Peter. "The Pooling and Tranching of Securities: A Model of Informed Intermediation," *Review of Financial Studies*, 18, 2005, pp. 1-35.

Gorton, Gary, and Nicholas Souleles. "Special Purpose Vehicles and Securitization," NBER Working Paper 11190, March 2005.

Gorton, Gary, and George Pennacchi. "Financial Intermediaries and Liquidity Creation," *Journal of Finance*, 45, 1990, pp. 49-71.

Klemperer, Paul. *Auctions: Theory and Practice*. Princeton, NJ: Princeton University Press, 2004.

Leland, Hayne, and David Pyle. "Informational Asymmetries, Financial Structure, and Financial Intermediation," *Journal of Finance*, 32, 1977, pp. 371-87.

Mei, Jianping and Anthony Saunders. "Have U.S. Financial Institutions' Real Estate Investments Exhibited 'Trend-Chasing' Behavior?" *Review of Economics and Statistics*, 79, 1997, pp. 248-58.

Milgrom, Paul. *Putting Auction Theory to Work*. Cambridge University Press, 2003.

Minton, Bernadette, Anthony Sanders, and Philip Strahan. "Securitization by Banks and Finance Companies: Efficient Financial Contracting or Regulatory Arbitrage?" Working Paper, Ohio State University, September 2004.

Modigliani, Franco, and Merton Miller. "The Cost of Capital, Corporation Finance and the Theory of Investment," *American Economic Review*, 48, 1958, pp. 261–97.

Myers, Stewart, and Nicholas Majluf. "Corporate Financing and Investment Decisions When Firms Have Information That Investors Do Not Have," *Journal of Financial Economics*, 13, 1984, pp. 187-221.

Plantin, Guillaume. "Tranching," Financial Markets Group Discussion Paper DP-449, Revised December 2004.

Do Budget Deficits Cause Inflation?

BY KEITH SILL

Is there a relationship between government budget deficits and inflation? The data show that some countries—usually less developed nations—with high inflation also have large budget deficits. Developed countries, however, show little evidence of a tie between deficit spending and inflation. In "Do Budget Deficits Cause Inflation?," Keith Sill states that the extent to which monetary policy is used to help balance the government's budget is the key to determining the effect of budget deficits on inflation. He examines the theory and evidence on the link between fiscal and monetary policy and, thus, between deficits and inflation.

In 2004, the federal budget deficit stood at $412 billion and reached 4.5 percent of gross domestic product (GDP). Though not at a record level, the deficit as a fraction of GDP is now the largest since the early 1980s. Moreover, the recent swing from surplus to deficit is the largest since the end of World War II (Figure 1). The flip side of deficit spending is that the amount of government debt outstanding rises: The government must borrow to finance the excess of its spending over its receipts. For the U.S. economy, the amount of federal debt held by the public as a fraction of GDP has been rising since the early 1970s. It now stands at a little over 37 percent of GDP (Figure 2).

Keith Sill is a senior economist in the Research Department of the Philadelphia Fed.

For a long time, economists and policymakers have worried about the relationship between government budget deficits and inflation. These worries stem from the possibility that the government will finance its deficits by borrowing or by printing money. Should deficit spending and a large public debt be worrisome for monetary policymakers who are concerned about the economy's level of inflation? Do government budget deficits lead to higher inflation? When looking at data across countries, the answer is: it depends. Some countries with high inflation also have large government budget deficits. This suggests a link between budget deficits and inflation. Yet for developed countries, such as the U.S., which tend to have relatively low inflation, there is little evidence of a tie between deficit spending and inflation. Why is it that budget deficits are associated with high inflation in some countries but not in others?

The key to understanding the relationship between government budget deficits and inflation is the recognition that government deficit spending is linked to the quantity of money circulating in the economy through the *government budget constraint*, which is the relationship between resources and spending. At its most basic level, the budget constraint shows that money spent has to come from somewhere: in the case of local and national governments, from taxes or borrowing. But national governments can also use monetary policy to help finance the government's deficit.

The extent to which monetary policy is used to help balance the government's budget is the key to determining the effect of budget deficits on inflation. In this article, we will examine theory and evidence on the link between fiscal and monetary policy and, thus, between deficits and inflation.

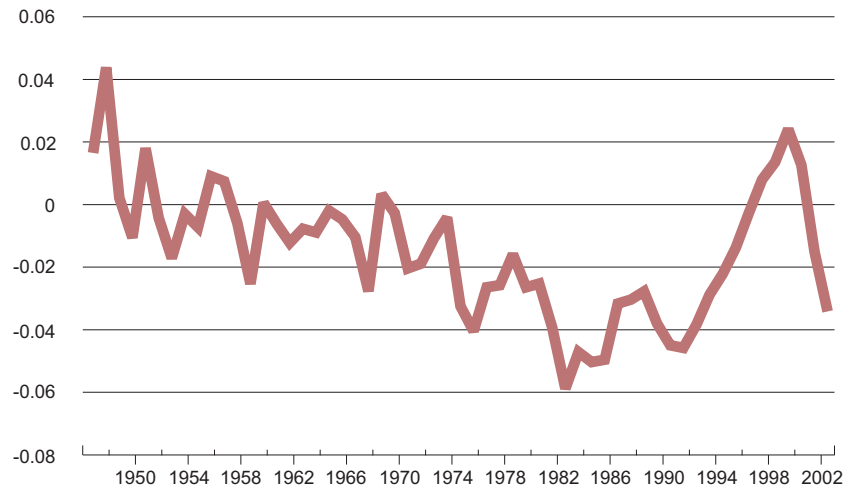
BUDGETS AND ACCOUNTING

Budget constraints are a fact of life we all face. We're told we can't spend more than we have or more than we can borrow. In that sense, budget constraints always hold: They reflect the fact that when we make decisions, we must recognize we have limited resources.

An example can help fix the idea. Imagine a household that gets income

FIGURE 1

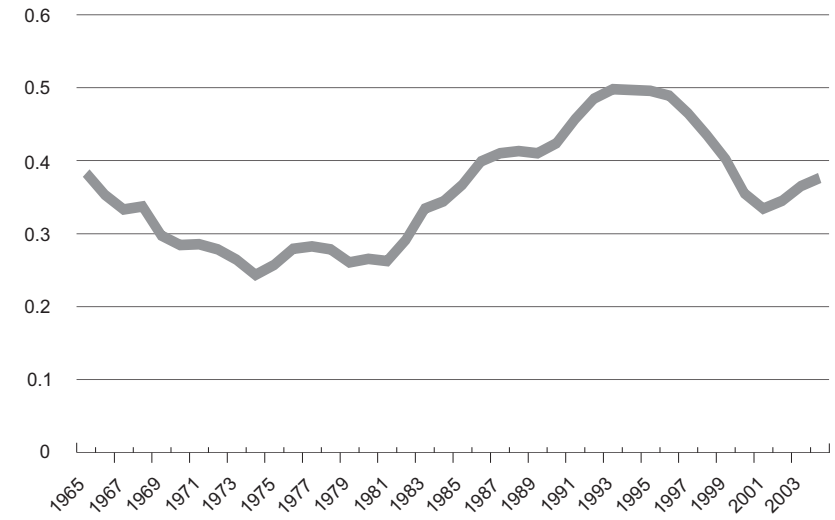
Federal Surplus/Deficit Relative to GDP



Source: Haver Analytics

FIGURE 2

Federal Public Debt Outstanding as a Fraction of GDP



Source: Office of Management and Budget, Flow of Funds Accounts

from working and from past investments in financial assets. The household can also borrow, perhaps by using a credit card or getting a home-equity loan. The household can then spend the funds obtained from these sources to buy goods and services, such as food, clothing, and haircuts. It can also use the funds to pay back some of its past borrowing and to invest in financial assets such as stocks and bonds.[1]

The household's budget constraint says that the sum of its income from working, from financial assets, and from what it borrows must equal its spending plus debt repayment, plus new investment in financial assets. There are no financial leaks in the budget constraint: The household's sources of funds are all accounted for, its spending is all accounted for, and the two must be equal. The household may use borrowing to spend more than it earns, but that source of funding is accounted for in the budget constraint. If the household has hit its borrowing limit, fully drawn down its assets, and spent its work wages, it has nowhere else to turn for funds and would therefore be unable to finance additional spending.

Just like households, governments face constraints that relate spending to sources of funds. Governments can raise revenue by taxing their citizens, and they can borrow by issuing bonds to citizens and foreigners. In addition, governments may receive revenue from their central banks when new currency is issued. Governments spend their resources on such things as goods and services, transfer payments such as Social Security to its citizens, and repayment of existing debt. Central banks are a potential source of financing for government spending, since the revenue the government gets from the central bank can be used to finance spending in lieu of imposing taxes or issuing new bonds. For example, the U.S. Treasury received a little more than $22 billion from the Federal Reserve in 2003.[2]

Much of a central bank's revenue comes from its monetary policy opera-

[1] The household can also sell some of its assets to finance consumption. This is tantamount to negative investment in assets.

[2] Recent detail on Federal Reserve payments to the Treasury can be found in the 90th *Annual Report*, Board of Governors of the Federal Reserve System, 2003, Table 5, page 270.

tions. An important aspect of modern monetary policymaking is controlling the short-term interest rate. Central banks do this by purchasing and selling interest-earning government bonds. If the central bank wants to raise the interest rate, it sells government bonds. If it wants to lower the interest rate, it buys government bonds. As a consequence of these *open market operations*, central banks have government bonds in their portfolios, and these bonds earn interest. Thus, one component of central bank revenue is interest earned on the government bonds it holds.

The second component of central bank revenue is also related to open market operations. Central banks are able to create and issue money to pay for the government bonds they purchase. The money that central banks create is called *high-powered money*, and it takes the form of currency held by the nonbank public plus the reserves banks are required to hold against certain types of deposits. Since the central bank can issue high-powered money to pay for things like government bonds, an increase in high-powered money represents a source of central bank revenue.

Revenues are one side of the central bank's budget constraint. What does the central bank spend its revenue on? As mentioned, a major use of funds is to purchase government debt in the conduct of open market operations. The other component of central bank spending is residual: what is left over after the central bank pays its expenses. In the U.S., this residual gets turned over to the Treasury each year.

We can get a *consolidated government budget constraint* by combining the budget constraints of the treasury and the central bank. The government spends its revenue on:

- Goods and services;
- Transfer payments; and
- Interest payments on government debt held by the public.[3]

This spending is funded by:

- Tax receipts;
- The increase in debt held by the public; and
- The increase in high-powered money.

Note that if the government increases the quantity of high-powered money it can reduce other taxes or borrowing.

The revenue the government gets from the increase in high-powered money is called seigniorage.[4] The extent to which governments use seigniorage as a means for financing budget deficits plays a key role in the link between budget deficits and inflation. Since the creation of high-powered money, and thus seigniorage, is undertaken by the central bank, the consolidated budget constraint shows the link between fiscal policy and monetary policy. Money creation is a source of revenue for the government. The amount of revenue the government gets from seigniorage has implications for the government's choices about taxes, borrowing, and spending.[5]

HOW MUCH CAN THE GOVERNMENT BORROW?

The consolidated budget constraint shows the link between the

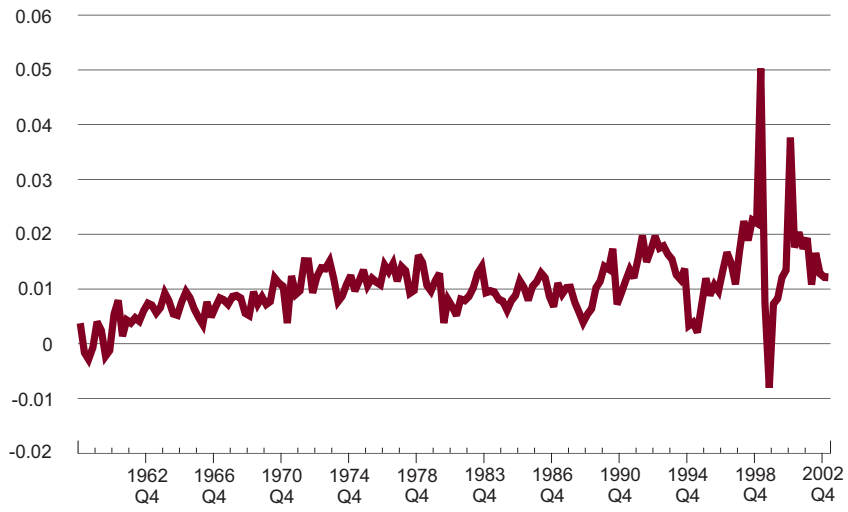
[3] Recall that interest paid on government debt held by the central bank goes back to the treasury.

[4] More technically, seigniorage is the real increase in the stock of high-powered money (currency held by the nonbank public plus bank reserves), i.e., the increase in the stock of high-powered money adjusted for the level of prices in the economy. As shown in Figure 3, for the U.S., this measure of seigniorage has been small. See the book by Frederic Mishkin.

[5] There is also a subtle way in which governments can use monetary policy to help finance spending. If the government can generate surprise inflation, the real value of the payments it makes to holders of its debt falls below what investors expected to receive when they bought the debt. Surprise inflation erodes the value of government debt, which means that a lesser amount of real tax revenue must be raised to pay off bondholders. However, generating surprise inflation to finance spending is ultimately a losing game for the government. Eventually, investors will catch on to what the government is doing and demand a high enough interest payment to compensate them adequately for the government's inflation policy.

FIGURE 3

Seigniorage Relative to Government Spending



Source: Haver Analytics

government's choices about spending, taxing, borrowing, and seigniorage. This relationship is a constraint only in the sense that there may be limits on the government's ability to borrow or raise taxes. Obviously, if there were no such limits, there would be no constraint on how much the government could spend at any point in time.

Certainly governments are limited in their ability to tax citizens. (That is, the government can't tax more than 100 percent of income.) But are governments constrained in their ability to borrow? Indeed they are. Informally, the value of government debt outstanding today cannot be more than the value of the resources the government has to pay off the debt.[6]

How do governments pay their current debt obligations? One way is for the government to collect more tax revenue than it spends. In this case, the surplus can be used to pay bond holders. Another way to finance existing debt is to collect seigniorage revenue and use that to pay bond holders. Finally, the government can borrow more from the public to pay existing debt holders. If the government chooses this last option, any new debt it issues would, in turn, have to be paid off using future surpluses, future seigniorage, or future borrowing. As long as the amount of debt the government issues to pay its obligations does not grow too fast over time, we can think of the current value of outstanding government debt as being ultimately backed by a stream of future surpluses and future seigniorage.[7] Since investors generally prefer to receive payouts sooner rather than later, the future stream of surpluses and seigniorage that backs government debt must be discounted to take account of the time value of money. That is, the current value of debt must equal the present discounted value of future surpluses and future seigniorage.[8]

Monetary policy does not necessarily have to adjust money growth in response to deficit spending... provided that deficit spending is expected to be offset by future surpluses.

We call this relationship the government's *intertemporal budget constraint*.[9] It indicates that the government must plan to raise enough revenue (in present value terms) through taxation and seigniorage to pay off its existing debt and to pay for its planned expenditures on goods, services, and transfer payments.

The intertemporal budget constraint has some interesting implications for monetary and fiscal policy. Suppose the government decides that, for a set path of future spending, it will lower current and future taxes permanently. This policy would lower the present discounted value of future surpluses. So, to fund the path of future spending, the government would need to increase the present discounted value of seigniorage. Since seigniorage is related to high-powered money growth, the implication is that money growth must increase in the future. Similarly, if the government decides to permanently increase future surpluses — for example, a permanent increase in taxes or a permanent reduction in borrowing — so that the present discounted value of future surpluses rises, the present discounted value of future seigniorage must fall; therefore, future money growth must fall.[10]

Note that the constraint does not say that an increase in deficits must be accompanied by a rise in seigniorage. An increase in the deficit could be temporary in the sense that it will be offset by future surpluses. In other words, a deficit today could be negated by a future surplus, so that the present discounted value of future surpluses remains unchanged. In that case, no offsetting adjustment in the value of discounted future seigniorage would be necessary. Monetary policy does not necessarily have to adjust money growth in response to deficit spending by the government, provided that deficit spending is expected to be offset by future surpluses. But if the present discounted value of future surpluses changes, there must be an offsetting change in the present discounted value of seigniorage, and vice versa.

POLICY, DEFICITS, AND INFLATION

Suppose that whenever there is a change in the present discounted value of seigniorage, fiscal policy adjusts so that the intertemporal budget constraint holds. In this case, monetary policy is independent in the sense that monetary policymakers take action without regard to fiscal policy, and then fiscal policy adjusts to maintain

[6] A formal derivation of this relationship can be found in the Technical Appendix.

[7] We have assumed that in the long run, government debt does not grow at a rate faster than the interest rate.

[8] Present value refers to an amount of money today that will become a given amount at a stated point in the future, depending on the interest rate. For example, if the interest rate is 10 percent, $100 today will be worth $110 in one year. So the present value of $110 one year from now (with an interest rate of 10 percent) is $100.

[9] An "intertemporal constraint" shows how government resources and spending are linked over time.

[10] The government can permanently increase future surpluses by raising taxes or borrowing less.

a balanced budget.[11] With monetary independence, policymakers are free to pursue goals such as low and stable inflation and not have to worry about using money growth to finance treasury budget deficits. In this case, we would not expect a tight link between government budget deficits and inflation because current government budget deficits are expected to be largely offset by future government budget surpluses. In addition, the path of government budget surpluses is expected to offset changes in seigniorage, so that the intertemporal budget constraint holds.

This does not mean that we may not observe some correlation between deficits and spending. For example, if the economy is hit by a recession, the deficit is likely to rise because tax revenues fall. At the same time, monetary policymakers may lower interest rates to combat the recession, an act that may subsequently lead to higher inflation. In this case, though, deficits are not, per se, the cause of inflation. Rather, deficits and inflation are both consequences of the recession.

The alternative case is one in which monetary policy is dependent. When monetary policy is dependent, the central bank adjusts seigniorage so that the budget constraint holds. Monetary policy responds to fiscal policy, so that seigniorage revenue becomes an important component of government finance. An independent treasury might decide to run permanent deficits, a situation that requires seigniorage to make up the gap between the value of the public debt and the present discounted value of budget surpluses. In this case, we could expect to see a link between deficits and inflation, since monetary policymakers respond directly to a fiscal policy of deficit spending. Whether monetary policy is independent and fiscal policy is dependent or vice versa is the key to answering the question of whether budget deficits imply higher inflation.

Dependent Monetary Policy May Result in Unexpected Outcomes. In a 1981 article, Thomas Sargent and Neil Wallace offer a famous example of how dependent monetary policy can lead to unexpected outcomes. Suppose fiscal policy is independent, monetary policy is dependent, monetary policy responds to fiscal policy, and the intertemporal budget constraint holds. In this case, an attempt by monetary policymakers to rein in inflation today by lowering money growth can result in higher inflation in the future: Policymakers are ultimately defeated in their efforts to lower inflation. How could this happen?

> It seems safe to say that, for the U.S. economy, there is little, if any, link between deficits and inflation.

Suppose monetary policymakers lower current money growth in an effort to bring down inflation. Lower money growth means lower seigniorage. If government spending and taxes do not change, the government will have to borrow more from the public in order to make up for the lost revenue from seigniorage. If the outstanding public debt increases, the intertemporal budget constraint implies that there must be a corresponding increase in the present discounted value of future budget surpluses and seigniorage. In a regime of fiscal independence, fiscal policy does not adjust, so the present discounted value of budget surpluses does not change. But that means that the present discounted value of seigniorage must rise to match the increase in the value of public debt outstanding. That is, the central bank will be required to increase the rate of money growth (seigniorage), an action that ultimately leads to higher inflation.[12] In this case, efforts to use monetary policy to lower inflation are self-defeating.

EMPIRICAL EVIDENCE ON INFLATION AND DEFICITS

Economic theory suggests that the strength of the relationship between government budget deficits and inflation depends on whether monetary policy is independent or dependent relative to fiscal policy. In countries where seigniorage is an important component of government finance, we are likely to find that government budget deficits and inflation are empirically linked. In countries with independent monetary authorities, the link between deficits and inflation is likely to be weaker.

Evidence for the U.S. Economy. As we can see from a plot of deficits and inflation for the U.S. economy since the end of World War II, there does not appear to be much of a relationship between government budget deficits and inflation (Figure 4). The contemporaneous correlation between federal budget deficits and inflation (GDP deflator inflation) is essentially zero. It is possible that deficits today are more highly correlated with future inflation than with current inflation — it may take some time for deficits to be felt in the form of higher inflation. But even if we look for the largest correlation between current deficits and future inflation, we find that it is still rather low at 10 percent, when current deficits are correlated against inflation

[11] See Michael Dotsey's article for more on independent and dependent monetary policy.

[12] There is a strong empirical link between money growth and inflation for a wide range of countries over a long span of time. See the article by George McCandless and Warren Weber.

six quarters ahead. It seems to be the case that for the U.S. economy, deficits and inflation are largely unrelated.

It seems safe to say that, for the U.S. economy, there is little, if any, link between deficits and inflation. The reason is that the Federal Reserve largely sets monetary policy independently of what the Treasury is doing to finance the federal government budget deficit. The Fed turns over its profit to the Treasury each year, but the Fed does not conduct monetary policy to raise revenue for the Treasury. Rather, the Fed focuses on stabilizing inflation and unemployment and does not conduct monetary policy with an eye toward financing fiscal deficits.

More thorough evidence than simple correlations bears out the finding that deficits and inflation are weakly linked, if at all, in the U.S. and, for that matter, in most of the world's advanced economies.[13] However, there does seem to be a link between deficits and inflation in the world's less-developed economies. For those countries, high inflation is often associated with high average government budget deficits.

Evidence for the Rest of the World. A recent study by Stanley Fischer, Ratna Sahay, and Carlos Vegh classified a sample of 94 countries into high-inflation and low-inflation countries. High-inflation countries, of which there were 24 in their sample, are those that experienced at least one episode of 12-month inflation exceeding 100 percent over the span 1960 to 1995. On average, inflation in those countries was a bit over 150 percent per year. Seigniorage as a fraction of GDP averaged about 4 percent in high-inflation countries versus an average of 1.5 percent in low-inflation countries. High-inflation countries rely more on seigniorage to help finance government spending. The authors find that for high-inflation countries, a worsening fiscal balance is much more likely to be accompanied by an increase in seigniorage than is the case in low-inflation countries.

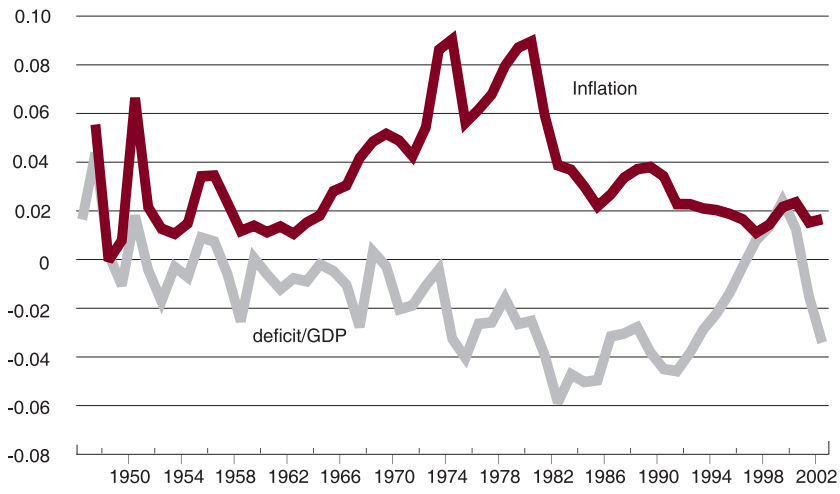
What triggers inflation? The authors use standard techniques to show that fiscal deficits lead to high inflation when the government depends on revenue from seigniorage to finance debt. They find that for high-inflation countries, a 10-percentage-point reduction in the fiscal balance (i.e., deficit) as a fraction of GDP is associated with, on average, a 4.2 percent increase in seigniorage. For low-inflation countries, there is no significant link between deficits and seigniorage. Also, when high-inflation countries experience episodes of low inflation, the link between deficits and inflation weakens dramatically.

A 2003 study by Luis Catao and Marco Terrones uses a broader sample of 107 countries over the period 1960 to 2001 to look for a link between fiscal deficits and inflation. They find a strong link between fiscal deficits and inflation in developing countries. For example, a 1 percent reduction in the ratio of the budget deficit to GDP is associated with an 8.75 percent lower inflation rate. Catao and Terrones also find results similar to those of Fischer, Sahay, and Vegh when the sample is broken into high-inflation and low-inflation countries using the 100 percent annual inflation rule. But they also find a statistically significant relationship between deficits and inflation in countries with moderate inflation as well, though the link is weaker. For low-inflation and advanced countries, Catao and Terrones find no link between fiscal deficits and inflation.

For developing countries, seigniorage is a significant source of revenue, and fiscal policy appears to be an important ingredient for the amount of inflation. Indeed, over the period

FIGURE 4

Federal Deficit and Inflation



Source: Haver Analytics

[13] An older set of empirical studies tended to find that there was at best a tenuous link between deficits and inflation for the U.S. economy. See the papers by G. Demopoulos, G. Katsimbris, and S. Miller; K. Grier and H. Neiman; D. Joines; and Robert King and Charles Plosser.

1980 to 1995, seigniorage as a fraction of GDP averaged about 2.2 percent, compared with only 0.64 percent in advanced economies such as the U.S., Germany, and Japan.[14] One possible reason for the greater reliance on seigniorage revenue in developing economies is that, for them, seigniorage may be a relatively efficient method to raise revenue compared with other forms of taxes. In developing countries, it may be difficult to collect tax revenue, since the tax base tends to be small and difficult to identify, especially when the government does not have a lot of resources to devote to building an efficient tax-collection system.

SUMMARY

Monetary policy and fiscal policy are linked because money growth, in the form of seigniorage, provides revenue to the fiscal branch of the government. But whether deficits lead to inflation depends on the extent to which monetary policy is independent, that is, the extent to which monetary policymakers must react to fiscal financing developments when setting policy goals and implementing them.[15]

For the U.S. economy, there is little evidence of a link between fiscal deficits and inflation, precisely because monetary policymakers have been free to pursue goals such as low and stable inflation. They are able to do this because fiscal policy is seen as sustainable, in the sense that deficit spending today is not expected to continue to the extent that monetary policy will have to provide major funding for the Treasury. This is largely the case for the developed countries of the world. Developing countries, however, often require revenue from seigniorage to meet their fiscal financing needs. Thus, these countries tend to show a strong link between fiscal deficits and subsequent inflation.

[14] For more detail on seigniorage revenue in developing and advanced economies, see the article by Paul Masson, Miguel Savastano, and Sunil Sharma.

[15] We have focused on the possible inflation consequences of government budget deficits. Other questions of interest we have not explored include the impact of budget deficits on real interest rates and exchange rates.

Technical Appendix

The Government's Intertemporal Budget Constraint

We can express the consolidated budget constraint in the symbolic form:

$$i_{t-1}B_{t-1}+G_t = T_t+(B_t - B_{t-1})+(H_t - H_{t-1})$$

where G_t is government spending at time t, $i_{t-1}B_{t-1}$ is interest payments on publicly held government debt outstanding, T_t is tax receipts, and H_t is high-powered money. The left-hand side of the expression is total spending by the government and the right-hand side is total sources of revenue. It is convenient to put the budget constraint in inflation-adjusted, or real, terms by dividing through by the price level P_t. Define the real interest factor as

$$(1+r_t) = \frac{1+i_t}{(P_t / P_{t-1})}$$

We'll use lower case to denote real values. Then re-arranging terms, we can write the consolidated budget constraint as:

$$(1+r)b_{t-1}+g_t = t_t+b_t+s_t$$

In this expression, t_t is the real value of taxes collected, and s_t is the real value of the increase in money, or seigniorage. Finally, $(1+r)$ is the real interest factor on government debt, which we assume (for simplicity) is constant over time. If we iterate the budget constraint forward T times into the future, we get:

$$(1+r)b_{t-1} = \sum_{i=0}^{T}\frac{t_{t+i}-g_{t+i}}{(1+r)^i}+\sum_{i=0}^{T}\frac{s_{t+i}}{(1+r)^i} + \frac{b_{t+T}}{(1+r)^T}$$

As long as the real amount of debt outstanding grows no faster than the real interest rate, which is a condition that says enough economic resources will be available to fully pay off any debt outstanding, then as T gets larger, the last term in the expression should get closer and closer to zero.

The first term on the left-hand side of the equal sign is the present discounted value of future budget surpluses. The second term is the present discounted value of future seigniorage. The equation shows that the real value of debt held by the public (principal and interest) is constrained by the government's ability to raise revenue to pay it off.

REFERENCES

Board of Governors of the Federal Reserve System. *90th Annual Report*, 2003.

Catao, Luis, and Marco Terrones. "Fiscal Deficits and Inflation," IMF Working Paper WP/03/65 (2003).

Demopoulos, G, G. Katsimbris, and S. Miller. "Monetary Policy and Central-Bank Financing of Government Budget Deficits," *European Economic Review*, 31(5), July 1987, pp. 1023-50.

Dotsey, Michael. "Some Not-So-Unpleasant Arithmetic," Federal Reserve Bank of Richmond *Economic Quarterly*, 82, 4 Fall 1996.

Dwyer, Gerald P., Jr., and R.W. Hafer. "The Federal Government's Budget Surplus: Cause for Celebration," Federal Reserve Bank of Atlanta *Economic Review*, July 1998.

Fischer, Stanley, Ratna Sahay, and Carlos Vegh. "Modern Hyper- and High Inflations," *Journal of Economic Literature* 60 (2002), pp. 837-80.

Grier, K., and H. Neiman. "Deficits, Politics and Money Growth," *Economic Inquiry*, 25(2), April 1987, pp. 201-14.

Joines, D. "Deficits and Money Growth in the United States 1872-1983," *Journal of Monetary Economics*, 16(3), November 1985, pp. 329-51.

King, Robert, and Charles Plosser. "Money, Deficits, and Inflation," Carnegie-Rochester Conference Series on Public Policy, 22, Spring 1985, pp. 147-96.

Klein, Martin, and Manfred Neumann. "Seigniorage: What Is It and Who Gets It?" *Weltwirtschaftliches Archiv.*, 126(2), 1990, pp. 205-21.

Masson, Paul R., Miguel A. Savastano, and Sunil Sharma. "Can Inflation Targetting Be a Framework for Monetary Policy in Developing Countries?" *IMF Finance and Development*, March 1998, pp. 34-37.

McCandless, George, and Warren Weber. "Some Monetary Facts," Federal Reserve Bank of Minneapolis *Quarterly Review*, 19(3), Summer 1995, pp. 2-11.

Mishkin, Frederic A. *The Economics of Money, Banking and Financial Markets*, 6th edition. Reading, MA: Addison-Wesley, 2001.

Sargent, Thomas, and Neil Wallace. "Some Unpleasant Monetarist Arithmetic," Federal Reserve Bank of Minneapolis *Quarterly Review*, 5(3), Winter 1981, pp. 1-17.

Walsh, Carl. *Monetary Theory and Policy*. Cambridge, MA: MIT Press, 2003.

Challenges and Opportunities in a Global Economy:

Perspectives on Outsourcing, Exchange Rates, and Free Trade

A Summary of the 2004 Philadelphia Fed Policy Forum

BY LORETTA J. MESTER

"Challenges and Opportunities in a Global Economy: Perspectives on Outsourcing, Exchange Rates, and Free Trade" was the topic of our fourth annual Philadelphia Fed Policy Forum held on December 3, 2004. This event, sponsored by the Bank's Research Department, brought together a group of highly respected academics, policymakers, and market economists, for discussion and debate about the macroeconomic impact of developments in the global economy. Our hope is that the 2004 Policy Forum serves as a catalyst for both greater understanding and further research on policymaking in an increasingly global economy.

Over the past couple of years, the widening U.S. trade deficit and rising oil prices became front page news in discussions of U.S. economic performance. The longer-term impact of globalization on our labor markets and economic well-being became a discussion topic at cocktail parties and around dinner tables. The feeling that globalization was leading to the loss of U.S. jobs made some people even question whether free trade was as positive for the U.S. economy as economists know it to be. As world economies become more integrated, topics such as the macroeconomic effects of outsourcing, exchange rate policies and the flow of financial capital, and free trade and the cross-border flow of goods and services are garnering increased attention from policymakers and researchers. How best to seize the opportunities and meet the challenges of the global economy was the focus of the 2004 Philadelphia Fed Policy Forum.

Loretta J. Mester is a senior vice president and director of research at the Federal Reserve Bank of Philadelphia.

Anthony M. Santomero, president of the Federal Reserve Bank of Philadelphia, began the day discussing the breadth and depth of the global economy's influence. The international marketplace is widening geographically, and the U.S.'s relationships with its traditional trading partners in North America and Europe, with Japan, and with the emerging markets of Asia are evolving.

In Santomero's view, developments in the global economy are transforming the basic structure of the economy, the issues policymakers need to address, and the questions researchers are studying. The revolution in information technology and the emergence of new market economies are opening up opportunities to reallocate production and distribution around the globe.

Yet, so far, the potential effects of this outsourcing on the U.S. economy have been difficult to quantify. Similarly, there is still much to learn about the distribution of the costs and benefits of free trade. An examination of the sharp decline in the value of the dollar during the mid-1980s suggests that a substantial relative price change causes an expansion or contraction of economic activity in well-established sectors but does not open up brand new areas of international trade. Declining trade barriers, however, bring more fundamental change to the economies affected. For example, as Timothy Kehoe discussed later in the day, the North American Free Trade Agreement (NAFTA) led to an increase in trade in goods and services that were traded only in limited quantities previously and accelerated the transfer of new technologies across borders. Santomero conjectures that one possible explanation for the difference in effects is that a change in tariffs is perceived as being more permanent than a change in exchange rates; hence, it elicits a larger response. He also posits another possible explanation: that changes in exchange rates

affect relative prices across a broader array of goods and services and so evoke smaller adjustments across that broad array, while changes in tariffs affect a smaller number of goods and services and so have narrower but larger effects.

While opening up free trade brings participants an improved standard of living, it also creates dislocations and imposes cost on individual sectors within nations. As Santomero points out, free trade is beneficial provided the people and firms who gain from it are able to compensate the losers. Policymakers need to grapple with the political problem of how to redistribute the benefits of free trade in order to build and maintain support for free-trade policy. Countries are approaching free trade along various paths. Some are pursuing global trade arrangements, others are pursuing free trade areas, and some are pursuing bilateral trade agreements. In Santomero's view, the success of each of these strategies in building the necessary support for free trade is an open question.

OUTSOURCING[1]

The Policy Forum's first session considered the issue of outsourcing. Was it a reason for employment's slow recovery in this expansion? What has it meant for the industrial sector? And what determines whether a firm will choose to outsource its operations?

Labor markets have been weaker for longer in this recovery than in any of the other postwar recoveries, even the one in 1991, which has been called the jobless recovery. The 1990 and 2001 recessions were about the same length – eight months – but it took almost four years for U.S. employment to recover back to the level of its previous peak in March 2001. During the 1991 recovery, it took about two and a half years.

Cathy Minehan, president of the Federal Reserve Bank of Boston, elaborated on the behavior of labor markets during this business cycle. In her view, foreign outsourcing has not played a major role in the relatively slow rate of job growth during this recovery. The U.S. economy in the third quarter of 2004 looked quite healthy, growing at a sustainable pace, with the unemployment rate trending down, inflation well contained, and productivity growth strong. Still, sluggish job growth had been a concern during the recovery. Labor-force growth had outpaced job growth during and after the recession and opened an employment gap. Unemployment had been longer in duration than typical, and Minehan posited that this was because job losses during the recession had been of a more permanent than temporary nature. Highly educated middle-aged workers lost jobs this time, but the less educated, younger workers made up more of the long-term unemployed. Also unique to this recovery is that labor-force participation continued to decline as the recovery unfolded.


Cathy Minehan

> Foreign outsourcing has not played a major role in the relatively slow rate of job growth during this recovery.

Minehan presented a range of estimates of how much job growth would be needed to close the gap between actual and full employment. These estimates depend on what is assumed about population growth, the labor-force participation rate, and the noninflationary unemployment rate. To meet demographic growth in the labor force, which includes population growth and changing patterns of work and aging, Minehan estimates the economy needs to add about 120,000 jobs per month. If labor force participation continues on the low side, then the economy needs to create fewer jobs to absorb labor supply. But if labor-force participation reverts to its more normal level, the economy would need to add more workers. Also, the lower one believes unemployment can go without inflation becoming a problem, the more jobs can be created. Depending on the assumptions about labor-force participation and the natural rate of unemployment, Minehan estimates that somewhere between 125,000 and 225,000 jobs per month would have to be created to absorb the increase in labor supply.

Minehan evaluated two factors that the media have often mentioned

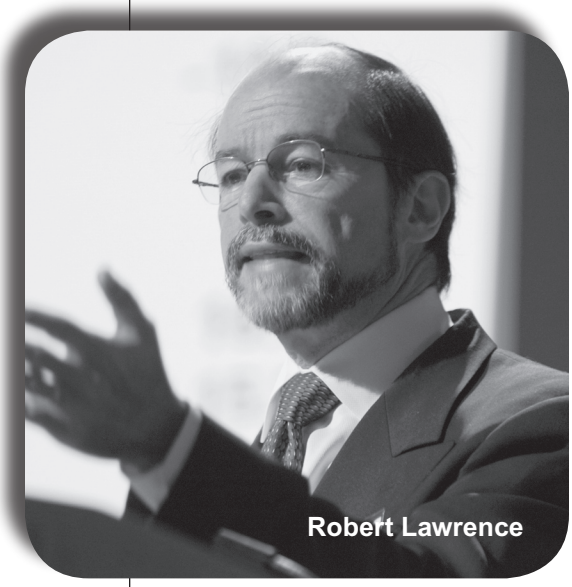
[1] Many of the presentations reviewed here are available on our web site at www.philadelphiafed.org/econ/conf/ policyforum2004.html.

as factors for the recent unusually slow job growth. First, the loss in manufacturing jobs has continued, and it has become steeper in recent years. But in Minehan's view, while this is part of the recent story, it cannot fully explain sluggish job growth, since the economy has been losing manufacturing jobs for most of the last 30 years. Second, foreign outsourcing has expanded. Not only goods-producing industries but also service-producing industries have begun to outsource. But, again, this cannot be the full explanation. While U.S. firms are outsourcing to India and China, Minehan points out that those countries appear to be buying more services from the U.S. than the U.S. is from them, and this creates an offset in terms of jobs. The fact that U.S. firms do not point to imports or outsoucing as the main cause of extended layoffs is taken by Minehan as evidence against the outsourcing explanation of slow job growth.

Then what is the explanation? Why are U.S. firms demanding less labor? Partly, this may be due to structural change as the economy shifts its mix of products and services; partly it might be a reaction to increasing labor costs, especially the cost of benefits; partly it might be firms' response to higher uncertainty, perhaps over the staying power of the recovery because of high oil prices and geopolitical concerns; and perhaps it's because firms are driven to become ever more productive. Minehan concludes that the latter two factors – uncertainty and the drive for increased productivity – might be the best explanations of the sluggish job growth that characterized this recovery.

Robert Lawrence of Harvard University extended the discussion of the relatively weak employment growth the U.S. experienced during the recovery. The media have focused on the role of international trade, particularly with China and India, and the effects of outsourcing on the U.S. economy were discussed during the recent presidential campaign. Lawrence described some of his recent research with Martin Baily of the Institute for International Economics that attempts to quantify the role of trade on the employment losses between 2000 and 2003. Like Minehan, he pointed out the sharp drop in manufacturing employment during the recent recession. In fact, while the share of employment in manufacturing declined throughout the 1990s, the number of workers employed in manufacturing didn't begin to decline until 2001, the beginning of the recession. In his view, one cannot simply attribute this to stronger productivity growth because productivity growth was rapid not only in manufacturing but also in other sectors, which experienced fewer losses. Looking deeper at the data shows that managers and production workers suffered the largest job losses, but many of those managers were in the manufacturing sector. Another factor during this recovery that Lawrence highlighted was the abnormally slow recovery in investment, which he feels is an important part of the story. Indeed, the largest manufacturing employment decline was in computers and electronic products, which lost about 30 percent of its jobs. In effect, it was the capital goods part of the manufacturing sector that experienced the highest job losses. In addition, exports during this cycle were quite a bit weaker than they were over other cycles, while imports were somewhat weaker. Since manufacturing productivity growth was much higher over the 2000-2003 period than either manufacturing export or import growth, jobs attributable to exports declined over this period, as did jobs embodied in imports.[2]

Devising reliable measures to determine trade's impact on employment is not easy.


Robert Lawrence

But devising reliable measures to determine trade's impact on employment is not easy. Lawrence and Baily take an input-output approach to determine the sectors in which exports create jobs and the sectors in which imports subtract from jobs in the sense that jobs in those sectors would have been higher had we produced those imports domestically rather than buy-

[2] Over the 2000-2003 period, manufacturing productivity growth rose 15.2 percent, manufacturing exports declined 8.8 percent, and manufacturing imports rose 2.3 percent.

ing them from abroad. Since total output equals production for domestic use plus exports minus imports, after jobs attributable to exports and imports are determined, jobs attributable to domestic use can be calculated as the residual. The results of their analysis suggest that weak U.S. domestic demand and trade both contributed to the loss in employment from 2000 to 2003, but that domestic demand had a larger effect than trade. Moreover, most of the job losses due to trade were due to weak exports and not to increased imports. Merchandise imports as a share of goods GDP were stable, 31.8 percent in 2000 and 31.4 percent in 2003, while merchandise exports as a share of goods GDP fell from 22.7 percent in 2000 to 20.1 percent in 2003. Based on data available as of December 2004, Lawrence and Baily estimate that of the 2.85 million jobs lost between 2000 and 2003, 2.54 million were due to weak domestic demand, 0.74 million were due to weak exports, and imports actually contributed 0.43 million jobs. Lawrence concludes that the job losses during the recession and first part of the recovery were "made in America." His analysis also reveals that the decline in U.S. exports is a market-share story rather than a weak-foreign-demand story. The U.S. lost competitiveness against other suppliers to the world market. If the U.S. had held its share in world markets, exports would have risen by 23.5 percent rather than declined. The lagged effects of the rise in the value of the dollar in the late 1990s played an important role in limiting U.S. exports as well.

Finally, Lawrence turned his focus to the future of manufacturing employment. Here there are two countervailing effects. If the U.S. closes the trade deficit by 2015, this will create jobs as U.S. exports increase and imports decrease. But if at the same time productivity growth in manufacturing stabilizes at its average 3.9 percent pace seen over the past decade, then net employment creation will be much less. Lawrence concluded that contrary to the discussion in the popular press, trade was not a large part of the story of the employment losses during the recession and it isn't likely to be a large part of the manufacturing employment story of the future.

While the session's first two speakers concentrated on the macroeconomic effects of trade and outsourcing, the next speaker, **Gene Grossman** of Princeton University, refocused the discussion, taking a microeconomics perspective on how multinational firms decide to organize their production activities. Grossman explained that trade theory is concerned with the allocation of resources over the longer run rather than the shorter-run dynamics discussed by our first two speakers. He began by explaining the difference between "outsourcing" and "offshoring," terms that are often used synonymously in popular discussions but that trade economists view as distinct. Outsourcing pertains to how a firm chooses to organize itself. Does the firm perform an activity in-house, or does it subcontract the activity to another producer? A decision to outsource is a decision to go outside the boundaries of the firm. Offshoring pertains to the location of an activity, either at home or abroad. A firm that subcontracts, say, its call center, to another firm that sets up the center in India, would be offshoring and outsourcing.

Outsourcing has several distinctive features. The types of goods that are traded in outsourcing relationships are often customized for a particular user. This is different from the types of products that trade theory usually considers, which are homogeneous goods that can be bought in multiple markets. This customization requires relationship-specific investments, which enhance the value of the relationship. Outsourcing also requires contracts to govern the relationship.

Offshoring also has distinctive features. One aspect is the cost of transportation and communication. These fixed costs can create complementarities between offshoring activities. Once one activity is moved offshore, it is cheaper to offshore another activity. This can lead to an increase in the volume of activity that is moved offshore. Thus, there's a positive feedback. Once a firm has paid the fixed costs of moving an activity offshore, say, to a low-wage country, the firm's unit costs of production will be lower and it will gain sales. But the increased sales give the firm the incentive to lower its unit costs in other ways, so the firm may consider paying the fixed costs to move another production activity offshore to achieve further reductions in unit costs. Also, if transportation costs are high, firms might move several parts of the production process offshore at the same time to economize on these costs. Thus, the economy can go from exhibiting a small amount of offshoring to exhibiting a large amount in a short period of time. Hence, the fact that U.S. firms aren't offshoring that much production yet does not imply that they won't in the future. Another aspect of offshoring is that it

It is often the largest and most productive firms that find it cost effective to move production offshore.

is often the largest and most productive firms that find it cost effective to move production offshore, since these firms are better able to bear the fixed costs needed to obtain savings on the variable costs of production and to bear the increased cost of monitoring performance across a longer distance.

The new literature on trade is drawing on the theory of the firm to address some of the interesting questions regarding outsourcing and offshoring. What accounts for the increasing fragmentation of the production process? What determines the form of offshoring? Does it differ by country? By industry? What characteristics of the firm or its activities help us understand the organizational mode it would choose?

With his co-author Elhanan Helpman of Harvard University, Grossman has studied some of the tradeoffs between outsourcing production versus producing in an integrated firm. On the one hand, specialized suppliers of inputs can usually produce more efficiently, especially if they provide those inputs to more than one customer. On the other hand, because not every contingency can be written into the contract, the supplier and the final producer may be subject to potential "hold-up" problems. The final producer may end up having to pay more than expected for the inputs. Or the supplier, after having made the relationship-specific investments needed to produce the specialized input, might find it difficult to get the final purchasers to share in the cost of those investments. This creates an incentive for underinvestment relative to what an integrated firm would do. Also, the supplier might do less customization of its input so that it could sell to other buyers if it has to. Once the input has been fully customized, it's harder to sell to any other buyer, and this puts the input supplier in a weak bargaining position relative to the buyer. Thus, the theory predicts that there would be a tendency toward less firm-specific investment and customization in industries with more outsourcing. In industries where these types of investment and customization are very important to the production process, integrated production rather than outsourcing would predominate.

The new literature on trade is drawing on the theory of the firm to address some of the interesting questions regarding outsourcing and offshoring.

Other research suggests that one mechanism for getting around the potential underinvestment and undercustomization problem between supplier and final producer is cost-sharing for the investments. We often see firms providing their suppliers with specialized equipment or lending them funds to purchase such equipment or raw materials. Cost-sharing on the labor side is much less common. But this cost-sharing means that the supplier has more bargaining power in any ex post renegotiations with the producer. This hold-up problem will be worse the more capital-intensive the production process is. Hence, this theory predicts that we would see more outsourcing in industries that are more labor intensive and less in industries that are capital intensive – and, analogously, more outsourcing to countries with abundant labor and less outsourcing to countries with abundant capital. This seems to fit reality.

Another feature of countries that firms would outsource to is what Grossman calls thick-market externalities. A firm is looking for a producer to customize its input, so it wants to find partners with the proper expertise to make what it wants. This could differ from what another producer is looking for. If we think of potential suppliers arrayed along a spectrum according to their type of expertise, finding someone with expertise close to what the producer is looking for is important. The denser or thicker the market of suppliers, the more likely the producer will find one with the expertise close to what he is looking for. There is a positive feedback. If more U.S. producers outsource business services to India, it will be more profitable for Indian firms to develop the expertise to provide those services. And as more Indian firms enter the market and develop the expertise, the easier it will be for a U.S. firm to find a suitable supplier in India. On the other hand, if no firms are outsourcing to a particular country, then a firm might not want to be the first to outsource there, since it might not find the expertise it is looking for.

Grossman's research also suggests that a country's legal environment is an important determinant of the volume of outsourcing the country can expect to obtain. An improved contracting environment, all else equal, makes the country more attractive to outsourcers. However, all else is not equal – eventually wages rise as the contracting environment improves, and this may lead firms to look elsewhere, especially if the original motivation for outsourcing was to save on labor costs.

Another tradeoff when considering outsourcing versus integration concerns the incentives the firm can give to managers for good performance. Since an external supplier has

to put up the cost of the inputs and the labor for producing the inputs, it typically has more at stake than an internal manager does, and this would provide a better incentive for good performance. On the other hand, it is probably easier for a firm to monitor the performance of one of its own internal divisions than an external supplier. These considerations imply that outsourcing will more likely be chosen by firms with very high or very low potential productivity and that firms with intermediate productivity will choose integration. In addition, for those firms that remain integrated, offshoring is chosen most often by the more productive of these firms. A look at the data suggests this seems to accord well with actual experience. However, economists are just beginning to empirically test the theories explaining firms' choices of outsourcing versus integration and home versus offshore production. According to Grossman, this empirical work shows promise, and it, along with new theoretical models, is helping us understand which types of firms in an industry are the ones that go offshore or engage in outsourcing, which types of industries are prone to these types of trade relationships, and in which types of countries we should expect to see one form of production versus another.

EXCHANGE RATES

The Policy Forum's next session looked at implications of exchange rate policies and trade deficits on the macroeconomy. **Jeremy Siegel** of the Wharton School, University of Pennsylvania, began the session emphasizing the demographic component of structural trade deficits or surpluses across countries, which in his view is often neglected. Over the past 50 years, life expectancy has risen and retirement age has fallen. In 1950 in the U.S., the difference between the two was only 1.6 years and today it is 14.4 years – a large change. However, these trends cannot continue. In 1950 in the U.S., the number of workers per retiree was seven to one. Now it is five to one, but it is slated to decline to two and a half to one by 2050. And other countries, including Japan, Italy, Spain, and Greece, are aging more quickly than the U.S. In Japan, the number of workers per retiree approaches one to one by 2050, which means the workers have to produce not only for themselves but also transfer goods to the retirees. These trends imply that retirement age has to increase. To investigate the effects of these demographic trends, Siegel has built an economic model to study who in the world is going to produce the goods and who is going to buy the assets in the economy. In the model, income grows at the rate of productivity growth until a person retires and then it is zero, and consumption grows at the rate of productivity growth until a person retires and then it is flat. The outcome of the model is the equilibrium retirement age, assuming that Social Security taxes are fixed. The model suggests that by 2050 the retirement age in the U.S. has to increase to 73, which implies that the difference between life expectancy and retirement age narrows to 9.2 years. As Siegel points out, it isn't merely that people have to work longer because they are living longer – the retirement age has to increase almost twice as fast as projected life expectancy. Things are worse if life expectancy rises more than the conservative estimates Siegel uses in his model simulation.


Jeremy Siegel

Economists are just beginning to empirically test the theories explaining firms' choices of outsourcing versus integration and home versus offshore production.

What can help solve this "age wave" problem? Faster productivity growth can help the situation, but only modestly. That's because when productivity growth accelerates, wages go up, and when wages go up, retirement benefits go up. So there's not much help there. Immigration might help. But a half billion immigrants into the U.S. over the next 45 years would be needed to keep the retirement age in the mid 60s; that number is far higher than the current U.S. population of 294 million.

Siegel says the hope comes from the developing world, where 85 percent of the world's population lives and where the population is much younger

than the developed countries'. The developing world's age profile is about 50 years behind that of the developed world – for example, the distribution of population by age group in India today looks like that of Japan or the U.S. in 1950. The number of workers per retiree is projected to decline in India but only to four to one by 2050. According to Siegel's model, if the developing world can grow at 6 percent per year into the future, which is optimistic but not overly so given current experience, then the retirement age in the U.S. and other developed countries can stay roughly where it is today. If growth in the developing world is less, then retirement age in the U.S. and other developed countries will have to rise. But assuming that growth in the developing world is 6 percent, then it is the developing countries that produce the goods and buy the developed world's assets. Today, the developing countries own less than 10 percent of the world's capital, but the model simulations suggest that by 2050, they will own most of the world's capital and they will be producing most of its goods. The model implies that the developing countries will be running large trade surpluses, while the developed countries will be running increasingly large trade deficits. Because most of their populations will be retired, the developed countries will need to import goods for consumption, and they will sell off the assets they have been accumulating for many decades. These trade flows come out of the demographics; they are not structural imbalances.

What are the implications for exchange rates? In Siegel's model, the trade deficits in the U.S. and the other countries of the developed world are sustainable at current exchange rates – they are driven by the demographics. Thus, even though the U.S. trade deficits are very large, they do not cause a depreciation of the dollar. As long as foreigners want to acquire U.S. assets and Americans want to acquire foreign goods, the trade deficits won't put pressure on the dollar exchange rate. Given this, Siegel suggests that when we are trying to determine whether a particular trade deficit is sustainable, we shouldn't use a zero deficit as the basis of comparison but the structural deficit that will obtain in the long run because of large differences in demographics across countries.

Because most of their populations will be retired, the developed countries will need to import goods for consumption, and they will sell off the assets they have been accumulating for many decades.

Michael Mussa of the Institute for International Economics followed Siegel with an opposing view of the sustainability of the U.S. current account deficit and the path of the exchange value of the dollar. Acknowledging that a wide range of outcomes for both exchange rates and the deficit have been observed in the past, Mussa made a case for why, in his view, the dollar remained overvalued. In his view, it is difficult for the U.S. to borrow against many of its assets on a world market. For example, borrowing against our domestic human capital is not really feasible. In the U.S., U.S.-owned assets abroad used to exceed foreign-owned assets in the U.S. by about 25 percent of GDP. Now, it is the opposite – foreign-owned assets in the U.S. exceed U.S.-owned assets abroad by about $2.5 trillion, or 25 percent of GDP. Mussa points out that no industrial country has ever seen that ratio go above about 60 percent of GDP, but if nominal GDP growth continues at about 5 percent and the current account deficit remains at about 6 percent of GDP, net external liabilities as a percent of nominal GDP would rise to 120 percent of GDP, double what any industrial country has been able to achieve and sustain. While it's possible the U.S. could sustain such a high level, he thinks it is unlikely.[3] Sustaining a current account deficit of 2 to 3 percent of GDP over the next decade or longer would be feasible in Mussa's view, since there are many reasons that foreigners want to invest in the U.S.

What's needed to bring the current account deficit down from 6 percent of GDP to what, in his view, is a more sustainable level? Mussa cites two things: first, a switch in the pattern of world demand toward purchases of U.S.-produced goods and services and therefore away from rest-of-world goods and services; second, an adjustment in the level of spending relative to income both in the U.S. and abroad. For the U.S. that means reducing our spending; for the rest of the world that means increasing their demand relative to their income.

Mussa says that both the private sector and the government sector in the U.S. will need to change their behavior to effect these changes. The

[3] In contrast, Jeremy Siegel said in the session's question and answer period that he thinks it is quite likely that the U.S. will break the historical maximum of 60 percent of GDP.


Gertrude Tumpel-Gugerell

private sector needs to save more; the government needs to put its Social Security and Medicare budgets in order. If the dollar depreciates, it will help reduce the drag from lower government spending by improving the U.S.'s net export position. If public-sector expenditures are not controlled, then if the dollar depreciates substantially, in Mussa's view, the Fed will need to raise interest rates to curb overly expansionary effects of higher net exports on U.S. economic growth.

Mussa believes the rest of the world faces a more difficult situation than the U.S., since they need to get demand up. In Europe, he looks to the European Central Bank to use monetary policy as much as it can. In Japan, Mussa thinks there is not a good deal more that monetary policy can do in the short run. The developing countries of Asia, which have been resisting an exchange rate correction, will need to allow that to happen. In Mussa's view these countries' massive interventions to buy dollars in order to keep their currencies from appreciating must slow down. Their purchases of U.S. Treasury securities as investment need not stop, but Mussa feels that $100 billion or $200 billion fewer purchases per year over the next couple of years would be a welcome development.

Mussa concluded with his perspective on whether a strong dollar is good or bad for the U.S. When the dollar is strong, the U.S. gets paid high prices for the goods and services it produces and sells abroad, and it pays relatively low prices for the goods and services it purchases from the rest of the world. All else equal, that is a good thing. But, again, all else is not necessarily equal. If the value of the dollar is so high that demand for U.S. goods and services by the rest of the world falls so that the U.S. doesn't earn enough on what it sells abroad to afford what it buys from the rest of the world, the U.S. will have to borrow more to finance the gap. And we may want the value of the dollar to fall in order to restore equilibrium. As Mussa puts it, the goal is to have the strongest dollar consistent with maintaining a sustainable equilibrium position in our external payments position over time. In Mussa's view, that's a dollar that is a fair bit weaker than seen in 2001 through early 2002, and significantly weaker than the current value on a trade-weighted basis – perhaps not much weaker against the currencies of most of the other industrial countries but much weaker against the currencies of a number of emerging market economies.

Gertrude Tumpel-Gugerell, a member of the Executive Board of the European Central Bank, brought an international perspective to the discussion. She discussed the question of whether large swings in exchange rates matter for the real economy and what the appropriate monetary policy response to exchange rate swings is.

The consensus of both academic economists and policymakers is that exchange rate movements are difficult to predict and that random walk models generally predict as well as standard macroeconomic models. Since exchange rates are asset prices, they are strongly influenced by expectations, which are difficult to measure and to include in formal models. Tumpel-Gugerell points out the irony, then, in the many calls for policy responses every time the value of the currency moves markedly. She views this as a sign that exchange rate movements are seen as important, despite all the difficulties in understanding them.

The consensus of both academic economists and policymakers is that exchange rate movements are difficult to predict and that random walk models generally predict as well as standard macroeconomic models.

The international monetary system has generally evolved toward greater exchange rate flexibility between major currency pairs. She distinguishes several major phases in which this evolution has occurred. The period of the gold standard was one of fixed exchange rates and convertibility of currencies into gold. It ended with the advent of World War I. During the interwar periods, coun-

tries went progressively back to the gold standard with the goal of restoring fiscal discipline. But this broke down again after the Great Depression convinced many that the system was faulty, and there was a brief period of flexible exchange rates. The Bretton Woods Agreement in 1944 led to all the major industrialized countries pegging their currencies to the U.S. dollar. This lasted until 1971. Tumpel-Gugerell calls the 1970s the trial-and-error system, which led to more flexibility in the 1980s and the 1990s, and the creation of the single European currency. She characterizes the current system as one of flexible exchange rates among major currencies accompanied by international cooperation.

Within the current framework, how much do exchange rate movements matter for the economy? Tumpel-Gugerell distinguishes between effects taking place through the price-competitiveness channel and those associated with market uncertainty. Regarding the former channel, theory suggests that exchange rate movements will have less of an effect on closed economies than on small open economies. Research suggests that a persistent exchange rate movement can have a significant effect on prices and GDP in the euro area, but that the effect usually is seen with a lag. For example, firms can squeeze their profit margins or can attempt to hedge against adverse exchange rate movements, thus delaying the effect of a persistent move. The European integration process, including the introduction of the euro, has reduced instability generated by shocks to the exchange rate. In Tumpel-Gugerell's view this reduction in volatility should help to boost trade across the countries in Europe.

But according to Tumpel-Gugerell, the main way to limit undesirable exchange rate instability is for policymakers to focus on achieving and maintaining sound macroeconomic fundamentals. She believes that if monetary policymakers are committed to price stability, this will lead to exchange rate stability over the long run. Sound and sustainable fiscal policy will also play a role in achieving economic balance among the world's economies.

FREE TRADE

I had the pleasure of moderating our final session, which looked at free trade. For economists, free trade is not very controversial – it offers participants the benefit of an improved standard of living. But the recent negative discourse in the popular press has led to a more nuanced discussion of the benefits – more documentation of those benefits – as well as discussion of the dislocations and other costs of the transition to free trade.

Timothy Kehoe of the University of Minnesota discussed how free trade agreements have affected trade and capital flows across countries. There has been an expansion of regional trade agreements in both Europe and the Americas. The U.S. signed a trade agreement with Chile in 2003, negotiated the Central American Free Trade Agreement with a number of countries, and has been in negotiations with several South American countries. The European Union has been expanding as well.

Kehoe focused his talk on the lessons we've learned over the past 25 years from the economic integration that has taken place – from our empirical experience with integration and from economic models, from where the economic models have worked well in predicting the effects of integration, and from where they have failed and need improvement.

While opening up to free trade and investment may be an important ingredient for generating economic growth, it is not sufficient.

Kehoe's first lesson is that while opening up to free trade and investment may be an important ingredient for generating economic growth, it is not sufficient. The Mexican Apertura, or opening up of the country, which began in the late 1980s and led to the North America Free Trade Agreement (NAFTA), had a large impact on Mexico, generating large increases in foreign trade and investment. Mexico now exports almost twice as much as the rest of Latin America combined. But while it generated significant growth in exports, it did not generate much overall economic growth – at


Timothy Kehoe

least not until after the 1994-1995 crisis there.

Lesson two is that a free trade area such as NAFTA or the European Union is neither necessary nor sufficient for generating foreign trade and foreign investment. Chile has just negotiated a free trade agreement with the U.S.; yet after its economic crisis in 1981-1982, its exports surged and are now about 25 percent of its GDP. Its GDP growth also accelerated sharply, and the increase was not only export driven. In contrast, Greece joined the European Economic Community in 1980, yet its exports as a percent of GDP are still under 10 percent, and foreign investment in Greece is also very small.

Lesson three is that to get foreign investment, domestic institutions such as banks are important; protections of investors' rights are important; property rights – like bankruptcy laws – are important. Although the Mexican banking system was opened to foreign participation in 1995, it still is not functioning well in financing private investment, which is still low compared to other countries like Chile. Thus, signing a free trade agreement is not a guarantee of direct investment.

Kehoe's fourth lesson reiterates Siegel's point, namely, that demographic differences can be important determinants of international capital flows. Mexico's baby boom was much stronger than the U.S.'s, and Mexico today has many young people. The median age in Mexico is 20 compared to 34 in the U.S. Similarly, the other countries in Latin America are young. In contrast, the European integration is between rich, old, and aging countries and poor, old, and aging countries. These demographics will affect both trade and capital flows across countries.

But Kehoe's fifth lesson is that capital flows may be substitutes rather than complements of trade flows. When we look at the U.S., we see that the volume of trade flows between the U.S. and our NAFTA partners is much higher than between the U.S. and the European Union, while the volume of investment flows is much higher with Europe. Kehoe posits that this might be because the U.S. is afraid of further trade restrictions and protectionism if trade volumes increase in Europe.

Lesson six is that applied general equilibrium economic models of NAFTA's impact did a poor job of capturing the very significant increase in trade volumes in North America, and they did a poor job of identifying the sectors in which trade increased. For example, if we compare one of the best model's predictions of U.S. exports to Mexico in different industry sectors over 1988-1999 to the actual data on exports, we find a correlation of less than 1 percent. One reason the models performed poorly is that they were unable to capture a fact shown in Kehoe's research: that much of the expansion of trade took place in sectors where there was little or no trade before trade liberalization. Models that focus on the exchange rate will not capture this new-goods effect; it happens with changes in trade policy.

Lesson seven is that dynamic applied general equilibrium models can do a good job capturing the path of capital flows when a country opens itself up to foreign investment. Flows of capital into a relatively poor country that opens itself to foreign investment are accompanied by trade deficits and depreciations of the real exchange rate. These inflows eventually stop, the trade deficit becomes a surplus, and the currency appreciates. Kehoe points out that this happened after Spain joined the European Community in 1986 and capital started flowing into the country. In 1992, the process reversed, and while the Spanish government was caught off guard and called the outflow of capital a crisis, this is exactly what the model predicted would happen.

> Despite the dislocations and reallocations that have to be borne, the steady march of technology and economic adjustments have allowed us to reap higher per capita income across the decades.

Kehoe's eighth and final lesson is that signing a free trade agreement does not always mean an increase in free trade. It depends on the level of trade barriers and tariffs the country operated under to begin with. In Kehoe's view, Ecuador's signing a free trade agreement with the U.S. is a large step toward free trade, since there's a high level of tariffs and trade barriers there. For Latvia and Slovenia, joining the European Union will give them access to European markets, but it will increase the level of tariffs under which they currently operate and so will be a step away from free trade. Kehoe predicts they will find it difficult to import to non-EU countries.

Douglas Irwin of Dartmouth College elaborated on the evolving debate over free trade. He pointed out that the first debates over U.S. trade policy took place when the new Congress met at Congress Hall, just a few steps away from the Philadelphia Fed. James

Madison of Virginia introduced the first tariff bill on the floor of the U.S. House when the first Congress met in April 1789. It passed in July, but only after a lively debate. Indeed, trade policy has always been a controversial aspect of U.S. economic policy. Perhaps the main reason is that trade is associated with economic change and it affects the distribution of income within the country. This means that trade is likely to always elicit various opinions. Irwin points out that the same arguments against trade tend to recur time and time again and that the current complaints that the U.S. can't compete because of low wages abroad, that foreign countries are unfair traders, and that trade will damage the economy have all been heard before. Nonetheless, the debate on trade has shifted over time. In the 1970s one of the issues was that multinationals were draining America of capital, investing in foreign countries rather than at home. In the 1980s, the debate focused on Japan and its high-tech development. In the 1990s, NAFTA was the issue. Currently, outsourcing to China and India has moved to the forefront. Irwin's study of history suggests that these issues will pass and, by 2010, a new country or issue will emerge as the focus of the debate.

When economists are asked if trade is good for the U.S. economy, the answer is yes. Despite the dislocations and reallocations that have to be borne, the steady march of technology and economic adjustments have allowed us to reap higher per capita income across the decades. Irwin acknowledges that going through the adjustments can be painful, but stopping the dislocation and economic change would create many more problems. And even though fear of trade has been constant through our history, the U.S. has consistently over the past 30 or 40 years pursued an agenda of opening up markets and keeping the U.S. market open. Irwin points out that the U.S. has done this in two ways. It has negotiated with foreign trade partners in the context of the World Trade Organization, and it has negotiated a number of regional and bilateral trade agreements. There is some debate among economists about whether the bilateral agreements are better or worse than multilateral negotiations, but both are proceeding with increased momentum.

This raises the question: if there is so much fear of globalization, why is it proceeding apace and why have markets remained open? Irwin points to three factors that help explain why there hasn't been a great backlash against globalization. First, domestic industries that compete with our imports, such as shoes and apparel, have been losing their political importance. They have shrunk in size or, in some cases, have been totally wiped out. For example, in the mid 1960s we imported a third to a half of shoes consumed in the U.S.; now we import over 95 percent. Also, a number of industries that faced foreign competition, such as semiconductors and automobiles, have gone global. In the past, they have argued for trade protection. Now, they've undertaken foreign investments, have diversified their production across many countries, and import many goods themselves. A second factor is that many U.S. imports are intermediate goods. Their consumers are businesses, not households, and they are dependent on getting these imports to carry out their own production. Irwin points to two examples: Canadian softwood lumber and steel. The users of Canadian wood have made it much more difficult for the U.S. government to give protection to domestic producers. Steel consumers put pressure on the Bush administration against steel tariffs. A third factor that has worked against protectionism is the macroeconomic stability the U.S. has enjoyed over most of the postwar period. Economic growth helps ameliorate the pain associated with the economic dislocations that accompany increased trade and the opening of markets.

Irwin pointed to an example that illustrates that protectionism is increasingly being viewed as a poor policy option. The state of Indiana has considered legislation to ban state contracts going to firms that outsourced to other countries. Not outsourcing the processing of state unemployment claims would cost the taxpayers of Indiana $16 million that could otherwise be spent on public works such as roads or schools, tax cuts, or servicing the debt. This cost has been publicized, and this, plus the fact that these jobs are not currently in Indiana anyway, has led many to question the proposed legislation.

In Irwin's view it will be difficult for trade opponents to move the U.S. away from its current very low tariff position and its open market. Irwin ended his presentation saying he believes there will always be critics of free trade and they will need to be rebutted by those who have a stake in and support the system of open world trade.

In recent testimony before the Finance Committee of the U.S. Senate,

Recent research has shown that countries that are more open to trade have higher ratios of private-sector credit to GDP.

Chairman Greenspan expressed the view that it is essential that we not put "our future at risk with a step back into protectionism."[4]

How can we ensure that U.S. markets remain open? Or, as **Raghuram Rajan**, economic counselor and director of research at the International Monetary Fund asks, how can we build constituencies for free trade? First, as was pointed out earlier in the day by Kehoe, it's important to have well-functioning institutions and well-defined property rights to realize the benefits of free trade. Those benefits include stronger economic growth. But also, over the 20th century, countries that have become more open for trade have tended to have better developed financial markets, which in itself helps to foster growth. This is an example of a positive feedback – better institutions allow the benefits of free trade and free trade allows development of better institutions.

Why do such correlations exist between openness and financial development? One possibility is that free trade strengthens the domestic constituencies for financial sector reform. For example, industries that want to begin trading more will need to finance that trade and will exert pressure on financial markets to develop to meet their needs. Or industries that feel the competition from foreigners could push for improved financial markets to aid them in remaining competitive. Recent research has shown that countries that are more open to trade have higher ratios of private-sector credit to GDP, and that seems to come about because the constituencies that are pro-finance become more powerful after trade liberalization.

But Rajan argues that the direction of causality may run the other way as well. The development of financial markets may increase the power of constituencies in favor of free trade relative to those opposed. Trade liberalization creates winners and losers; it does not make everyone uniformly better off. So to understand how constituencies in favor of free trade are developed, one must identify the winners and losers. Economic theory suggests that those who have the endowments in which the country is rich will be more pro-trade, since those who are relatively higher endowed will benefit from trade. For example, the U.S has more highly educated people than other countries. So opening up U.S. markets to trade will tend to benefit these people, since the U.S. is the country that can supply this type of worker. Thus, they are the ones that are pro-trade in the U.S. The low-skill workers in the U.S. will be hurt by free trade, since other countries can supply low-skill workers. Thus, in the U.S., the low-skill workers will tend to be against free trade. In poorer countries, where low-skill workers predominate, the more highly educated tend to be against free trade.

The development of financial markets may increase the power of constituencies in favor of free trade relative to those opposed.

But if free trade is beneficial overall, why can't the winners compensate the losers? Rajan conjectures it is because many of the required side payments would need to be enormous, and they would have to take place over such a long period of time that they would be hard to commit to.

If this is the case, then how does a country go about changing the political balance in favor of free trade? Rajan sees three broad possibilities. The first is through committing to external agreements like those of the World Trade Organization or setting a date far in the future when the trade and capital markets will open up, for example, the United Kingdom's big bang. The second possibility is through a crisis, as happened in India. The crisis exposes the fact that the country's policy of closed markets creates very bad outcomes, or the crisis reduces the relative political power of the status quo, who are against open markets. The third possibility is through building constituencies. In developing countries this entails showing them that there is more opportunity. The more trade that is occurring outside a country's borders, the more its own firms want to partake. Also, when the rest of the world is enjoying more flows of goods and capital, there can be more leakages across a country's borders, and the country may find it more advantageous to open itself up and control the flows, rather than have them go on without any control. Consumers also see the benefits of free trade in the form of lower prices and can create a pro-trade constituency. Firms that are more efficient are less likely to fear the increased competitiveness that comes from opening up markets. Hence, increasing entry into their industry can create more efficient firms that then emerge as a free-trade constituency. Similarly, individuals may fear free trade because they don't have access to education or the resources that will enable them to handle the changes that free trade will bring. Creating a safety net for these individuals will help shift their opinions regarding free trade.

[4] See Testimony of Chairman Alan Greenspan on China, before the Committee on Finance, U.S. Senate, June 23, 2005, www.federalreserve.gov/boarddocs/testimony/2005/20050623/default.htm.

Rajan concluded his presentation with some data from the World Value Survey, a survey of over 150,000 individuals in 66 countries between 1981 and 2000, which shows that preferences for competition, a proxy for free trade, do vary with factors such as education, income, age, and type of occupation. It turns out that younger people are more against competition than older people. This might reflect the fact that younger people tend to be producers and fear the job loss and older people tend to be consumers and value the lower cost of goods. Those with higher wealth, higher social status, and higher education tend to favor competition. Unskilled workers are more against competition than moderately or higher skilled workers. An interesting finding is that small business owners' attitudes toward competition are influenced by their access to credit, while managers' and employees' attitudes are not. Small-business owners in countries with strong credit markets are much more likely to be pro-competition than those in countries with weak credit markets. That is, if they have access to resources and feel they can get the resources to run their businesses, they favor competition. This is evidence that institutions matter and that financial development and well-functioning institutions that allow access to resources can foster freer trade – the reverse causality mentioned earlier. It also suggests that a country that finds itself with dysfunctional institutions might find it very hard to build support for changing those institutions and to build a constituency for free trade. Can an institution like the International Monetary Fund help? In Rajan's view the answer is yes, but only at the margin. There needs to be momentum within the country itself for change. Large, developed economies can help develop that internal momentum by helping to ensure that trade spreads to the poor, developing countries. Freer trade offers outside opportunities to the people in those countries, who can then develop into a constituency within the country in favor of even more openness and freer trade.

SUMMARY

The 2004 Policy Forum generated lively discussion among the program speakers and audience participants on a number of the challenges and opportunities brought by an increasingly global economy. Our hope is that the ideas raised will spur further research and foster a greater understanding of today's economy.

We will hold our fifth annual Philadelphia Fed Policy Forum, "Fiscal Imbalance: Problems, Solutions, and Implications," on Friday, December 2, 2005. You will find the agenda on page 35.